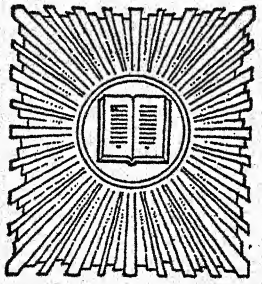
PREPARATION AND USE OF NEW-TYPE EXAMINATIONS

A MANUAL FOR TEACHERS

BY
DONALD G. PATERSON
Professor of Psychology
University of Minnesota



Yonkers-on-Hudson, New York
WORLD BOOK COMPANY
2126 Prairie Avenue, Chicago

WORLD BOOK COMPANY

THE HOUSE OF APPLIED KNOWLEDGE

Established 1905 by Caspar W. Hodgson

YONKERS-ON-HUDSON, NEW YORK

2126 PRAIRIE AVENUE, CHICAGO

The ultimate absurdity in examinations of the old type was reached in China, where young men seeking degrees in the classical learning of their country were locked in cells for weeks, while they toiled to exhaustion and sometimes to death over their endless assignments. Temperament, physical endurance, and memory were tested, it seems, rather than intellectual attainment. In Europe and in our own country, too, examinations have been a nightmare to students and to teachers from time immemorial. It is one thing to do the work prescribed for a given course, and it is another thing to pass the examination marking the completion of that course. It is one thing to teach, and it is another thing to devise examinations that will operate accurately and justly. Every teacher and every student is aware of the difficulties in the old system of giving examinations, yet examinations have seemed indispensable for groups of any size. Some measure of relief from outworn methods has been effected through the use, in our schools, of standardized intelligence tests prepared by psychologists specializing in the science of mental measurement; and the use of these tests has pointed the way to a wider application of the principles underlying them. The purpose of the present manual is to make the technique of preparing new-type examinations sufficiently clear to teachers to enable them to construct such examinations for ordinary classroom use, where standardized intelligence tests are not feasible

PPUNTE-5

PRINTED IN U.S.A.

CONTENTS

PREPARATION AND USE OF NEW–TYPE EXAMINATIONS

I

Introduction

Much has been published recently concerning the desirability of new-type examinations. This material, appearing both in book form and in various educational journals, deals for the most part with logical arguments and statistical evidence in favor of these objective examination methods. Considerable space is devoted also to the printing of sample copies of new-type examinations in various subjects, with a suggestion, here and there, concerning questions of procedure. But there is no one article, book, or manual containing in convenient form a summary of the best rules to be followed in preparing and using these examination methods. Believing that such a summary will render real service to the many teachers contemplating experiments with these methods in their own classes, the writer presents this manual. It describes the various forms of new-type examinations, lists the advantages and disadvantages of each, and makes suggestions for their preparation and use, in the form of twenty-two rules of procedure.

This manual does not aim to convince teachers that such examination methods should be used, nor does it present detailed evidence of their value. Its sole purpose is to aid those who, for one reason or another, desire to experiment with those methods. Arguments and evidence in favor of such methods or against them are contained in the various articles listed in the Bibliography (pages 79–87).

Without attempting to evaluate the success or failure of new-type examinations, it may be suggestive to list at this

time some of the subjects in which the new-type examination method has been used. The following list has been compiled from a survey of the literature dealing with this topic, and, while not complete, it is indicative of the range of subjects in which the method is applicable: Algebra, Animal Biology, Arithmetic, Chemistry, Civics, Civil Engineering, Civil Service Examinations, Contemporary Civilization, Dental Courses, Economics, Educational Psychology, English Literature, Examinations for Teacher's Licenses, French, General Science, Geography, Geometry, Government, Greek Art, History, Home Economics, Latin, Law, Medical Courses, "Orientation" Courses, Philosophy, Physics, Physiology, Political Science, Psychology, Rhetoric, and various vocational subjects such as Woodwork, Electricity, Sheet Metal Work, Auto Mechanics, and Bricklaying.

II

Definition of New-Type Examination

It is difficult to formulate a concise definition of what constitutes a new-type examination, since many of its features have been incorporated in traditional examinations. Perhaps an attempt to contrast the new type and the old type with respect to a few specific features will best serve our need for definition.

FORM OF QUESTION

In the first place, probably the outstanding characteristic of the old-type examination is to be found in the form of the question and the required response. The questions, very frequently, are of the "how" type, requiring a description of some process or the statement of the detailed logical steps of an involved explanation. For example, one of the questions in a recent physiology examination is, "Describe the secretion of gastric juice." This is a "how" type of question and could be stated without any alteration of meaning as, "How is gastric juice secreted?" The required answer, if as complete and adequate as that given by the textbook used in the course, demands the arrangement of forty-three items or ideas under twelve different headings. The new-type examination questions, on the other hand, are more frequently of the "what" type, requiring only a single-word answer. In testing a person's knowledge of a total process or his knowledge of the detailed logical steps of an involved explanation, the important critical steps are isolated and short-answer key questions are asked which, at least in many cases, presumably could not be answered by the student unless he possessed an adequate knowledge of the whole process.

The old-type question tests knowledge of a complicated process directly by requiring its production or application in relative detail, whereas the new-type question endeavors to test precisely the same thing indirectly by requiring short answers to critical key questions. Knowledge of the total process is thus inferred (hence measured indirectly) from the answers to a few critical key questions.

NUMBER OF QUESTIONS

In the second place, the old or traditional type of examination is generally composed of a relatively small number of questions. Sometimes the number is as low as five for an hour examination, and sometimes the number is as large as twenty or thirty (counting smaller subdivisions). The average number of questions is probably in the neighborhood of ten or twelve. The new type of examination, on the other hand, seldom contains fewer than fifty questions, the usual number varying from eighty to one hundred questions, and may even reach one hundred and fifty questions for a one-hour examination. It is possible to include as many as three hundred questions in a two-hour examination and still give adequate opportunity to practically every person for answering each question.

FORM OF ANSWERS

Thirdly, the pupil is required to write his answers to the old-type examination in one or more bluebooks or on sheets of theme paper, the questions being mimeographed on a sheet of paper or else written on a blackboard, whereas the pupil in taking the new-type examination is given the questions on mimeographed sheets and is required to record his answers either by writing a single word or two per question

PREFACE

THE writer, three years ago, presented to the faculty of the School of Business and later to other faculties in the University of Minnesota a brief paper on the improvement of the examination function in teaching. In these talks, an effort was made to present the claims of the new-type examinations, the writer offering his services as consultant to those desiring to experiment with these newer methods. He soon realized that a manual of directions for the preparation and use of such examining methods would be a decided aid to those contemplating their trial. Accordingly, such a manual as is here presented was planned, but it was delayed until circumstances permitted its final completion.

It is hoped that this manual will not only be welcomed by an increasing number of college and university instructors but will also be welcomed by high school and elementary school teachers as meeting the need for examining devices intermediate between the traditional examination and the more carefully devised and standardized educational achievement scales. The increasing use of short-answer examinations in elementary schools, high schools, and colleges warrants belief in the need for just such a manual as is here presented.

The adaptation of these objective examining methods by the United States Civil Service Commission, the Board of Examiners of the New York City Board of Education, and similar examining and licensing agencies forecasts a growing demand from various examining boards for help in incorporating these newer-type questions in their traditional examining procedures. It is hoped that this manual will meet also, in part, this demand.

The writer is indebted to Dr. J. B. Johnston, Dean of the College of Science, Literature, and the Arts, University of

Minnesota, for encouragement, assistance, and advice given in connection with the actual preparation of this manual. Sincere thanks are due also to Dr. Richard M. Elliott, Chairman of the Department of Psychology, for his painstaking criticisms and suggestions, which have immeasurably improved the first rough draft. Space does not permit adequate acknowledgment of the aid of others, but the writer desires to mention particularly his indebtedness to Dr. Charles Bird for helpful criticisms and to Dr. Arthur S. Otis, Test Editor of the World Book Company, who has contributed many worth-while criticisms and suggestions of real value in rounding out the manuscript in final form for printing. He is also under obligations to Dr. W. S. Foster and Dr. W. S. Miller for suggestions, and to Mr. J. E. Bohan for aid on the Bibliography. And, finally, grateful appreciation is extended to his wife for aid in the preparation of the original manuscript, in revising it, and in the careful reading of proof.

DONALD G. PATERSON

MINNEAPOLIS, MINNESOTA

or by making check marks to indicate correct answers. This contrast may be expressed fairly by saying that the old-type examination requires a maximum of writing, as opposed to a minimum in the new-type examination.

SUMMARIZED DEFINITION

By way of summary, we may characterize the old-type examination as requiring relatively long explanatory written answers to a small number of "how" type questions, whereas the new-type examination requires exceedingly short answers to a relatively large number of "key" questions, correct answers being symptomatic of total organized knowledge.

III

Principles Underlying Adequate Examinations

Before proceeding to describe in detail the forms of new-type examination questions, it is well to consider rather carefully a few elementary principles underlying adequate examinations.

1. Definition of Function to be Measured

The first principle to be considered involves the need for a definition of the capacities or abilities to be measured by the examination. Theoretically this means that an analysis of the objectives of each course be made and the examination be designed to measure the extent to which the objectives have been realized for each student examined. In the absence of scientific analyses of objectives in most courses, we may define the function to be measured in school examinations as knowledge of, and ability to think in, the subject matter of the various courses.[1] In the light of this definition the old-type examination does not suffer by comparison with the new-type examination perhaps as far as testing ability to think in the subject matter of the course is concerned, but it does seem to be deficient in testing subject-matter knowledge because its small number of questions sample only a small fraction of the knowledge which a thorough mastery of a course involves. Therefore the new-type examination can claim the honors with respect to the thoroughness with which it tests knowledge of a course.

[1] This definition is a modification of Ben D. Wood's definition of the function of college examinations as being "knowledge of, and ability to think in, the materials of the course," as given in his book, *Measurement in Higher Education* (World Book Company), page 153.

2. OBJECTIVITY IN MEASURING ACHIEVEMENT

The second principle involved in the production of adequate examinations has to do with objectivity. Objectivity in measurement implies the existence of measuring scales which when used by two or more competent examiners give precisely the same results. In other words, a measuring instrument is said to be objective when the factor of personal opinion or when the factor of the so-called personal equation is ruled out as completely as possible. It is evident that the use of clinical thermometers for the measurement of temperature has almost completely ruled out the personal equation which was formerly present when doctors merely estimated the degree of fever by feeling the hands and face of a patient. This is a concrete illustration of the replacement of subjective judgment by an objective measuring instrument, with a resultant greater uniformity and certainty for determining whether a person has a fever or not. A somewhat analogous situation prevails with respect to school examinations. The traditional essay examination tends to be subjectively evaluated, the personal opinion of the teacher playing an important rôle. In the new-type examination the scoring is much more objective.

This principle of objectivity requires a large number of small units, each scorable in definite units easily agreed to by all competent examiners. Furthermore, this principle requires that the examination involve relevant responses only, all irrelevant factors being eliminated. In both these respects the traditional examination is defective. Instead of a large number of small units it is composed of a small number of large units, and, what is more serious, these large units exhibit tremendous variability when one examination paper is compared with another. In addition, we must admit that the traditional examination does involve a host

of factors that are quite irrelevant as far as knowledge of, and ability to think in, the subject matter under consideration is concerned. What teacher will assert that his own judgment of what constitutes an adequate answer remains constant from paper to paper and from day to day? What teacher will affirm that he is really uninfluenced by a candidate's neatness and arrangement of the parts of the answers on the page, spelling ability, legibility of handwriting, composition ability, use of pet phrases of the lecturer or the text, etc.? To put the matter bluntly, the traditional examination may be likened to a battle in which the teacher is bombarded by, and sprayed with, innumerable words, phrases, sentences, and paragraphs by each student. The point at issue lies in the fact that much of this verbiage is irrelevant and inevitably results in a grading process which is highly subjective. The new-type examination corresponds more nearly with the requirements of the principle of objectivity. It is composed of small units requiring definite precise answers which are either right or wrong; hence any two teachers will agree in grading the same papers. Composition ability, spelling ability, legibility of handwriting, and neatness and arrangement of the parts of the answer are irrelevant factors and are ruled out because writing is reduced to a bare minimum. If such abilities are believed to be important in any given course, then they should be measured separately. Scientific method necessitates the measurement of one variable at a time; hence knowledge of, and ability to think in, the subject matter of the course should be measured objectively in the manner indicated and then separate examinations be devised to measure legibility of handwriting as such, another to measure spelling ability as such, and still another to measure composition ability as such. With respect to the principle of objectivity, then, it would seem that the new-type examinations score heavily,

because in this form of examination the personal opinion of the teacher is reduced to a minimum.

3. THE EXAMINATION SHOULD BE COMPREHENSIVE

The third principle underlying adequate examinations deals with what we might term the comprehensiveness of the examination. Every examination must of necessity attempt to measure but a part of the total knowledge possessed by the pupil examined. Total knowledge is inferred from the answers to questions which sample but a part of that total knowledge. This principle is the same as that which governs the work of the geologist in sampling the ore of a mine or of the grain dealer who samples the grain in a carload lot. *The sampling must be thorough and representative of every part of the thing sampled.* This principle is basic, and it is this very principle that is most flagrantly violated by the traditional examination. That is, it fails adequately to cover the subject matter of the course. Hence it is not equally fair to all, for a pupil may have prepared thoroughly on certain portions of the subject and be unlucky enough to have the questions bear heavily on other parts. This lack of adequate sampling is very likely to result in an unreliable score for any individual pupil. If one wishes to overcome this defect by making the traditional examination really comprehensive, a dilemma inevitably results because of the limitations of time allowed for the examination. To write discursive, explanatory answers to even a small number of questions which require elaborate descriptions of complicated processes is a time-consuming business. Not only is it time-consuming, but too much wasted effort is involved in proportion to the information extracted. Speed of writing, resistance to fatigue, and immunity from writer's cramp are all at too great a premium. The situation is reversed when we turn

our attention to the new-type examination. The questions are brief, the answers are short, and little or nothing is demanded by way of writing. Hence most of a pupil's time is consumed in cerebral activity or in thinking, and practically none of his time is devoted to grammatical construction, composition, or laborious handwriting. This makes it possible to multiply the number of questions and to sample thoroughly every important phase of the subject matter. For this reason the new-type examination may be made much more comprehensive and therefore more reliable, because chance influences are greatly reduced.

4. AN ADEQUATE EXAMINATION MUST HAVE HIGH RELIABILITY

In mental-measurement work reliability refers to the degree to which a test or examination measures whatever it purports to measure. The term refers to the consistency with which the examination measures the relative abilities of a group. The reliability or consistency of an examination may be determined by giving duplicate forms of the same examination to the same group to discover if each person receives about the same grade on both examinations. Or the odd items and the even items in a single examination are graded separately and the two scores for each person compared. Or the same papers are graded independently by two teachers and the two sets of grades compared. Or one paper is graded by a number of teachers and the agreements and disagreements in grades assigned to that paper are noted. When such tests of reliability are made of typical examinations, the results have proved startling.

A good summary of such reliability investigations is given in D. Starch's book, *Educational Psychology*. Results of one sample study will be given here. Marks given by 116

competent teachers to the same geometry examination paper varied from 28 to 92, the average mark being 69.9, standard deviation of the mean ± 1.2 and standard deviation of the distribution ± 12.8. Considering 70 as the passing mark, it is evident that approximately one half of the teachers would have passed the paper and one half failed the paper. Such a result is no reflection on the competency of the teacher; it is simply due to the defects inherent in any such subjective measuring scale. The causes of the unreliability of the traditional type of examination are many, arising chiefly because of (1) the inadequacy of the sampling, the questions being too few in number, (2) the complex nature of the required answers, (3) the lack of standardized scoring units, and (4) the presence of irrelevant factors influencing judgments of the real merit of the given answers. Now these weaknesses, inherent in the essay type of examination as ordinarily given, are mentioned, not for the purpose of slandering the efficiency of teachers, but rather to emphasize the complexities and difficulties of accurately measuring the achievements of pupils in school subjects.

The very factors that reduce the reliability of the traditional examination are largely overcome by the new-type examination; hence it is not strange that the work of Ben D. Wood, G. M. Ruch, and others has invariably shown the new-type examination to be much more reliable than the traditional examination as ordinarily prepared, given, and graded. Ruch's conclusions are so important as to justify quotation: "10 to 20 minute examinations of objective type are very much more reliable than 5 to 10 question traditional examinations which require 30 to 60 minutes." [1] Additional evidence on this point could easily be presented, but since the aim of this manual is not that of presenting a

[1] G. M. Ruch, *The Improvement of Written Examinations*, page 114. Scott, Foresman & Co., Chicago.

complete case in behalf of the new-type examinations, th writer is content to refer readers desiring additional data t the books mentioned here and to other similar pieces of wor mentioned in the Bibliography.

5. ECONOMY OF TIME AND EFFORT

The adequacy of an examination depends in part upo economy of time and effort and upon ease of administratio and scoring. Here, again, the new-type examination score heavily. There is economy of time on the part of the pupi in taking the examination, for he can give you ten times a much information in a given unit of time, or he can give yo as much information in one tenth the time now demande by the traditional essay examination. There is also a rea economy of effort on the part of the pupil, for he is freed from the dangers of writer's cramp and is freed from the laboriousness with which he has traditionally been forced to organize his lengthy answers in correct grammatical form. The writer does not know of an instance where pupils have not expressed a decided preference for the new-type examination. Furthermore, the majority of pupils have indicated that the new-type examination is more exacting with reference to the need for being fully prepared in the subject in order to make a good showing in the examination. Not only is there economy of time and effort on the part of the pupil, but there is an even greater economy on the part of the teachers. In fact, a great load can be lifted from th shoulders of the teachers by the adoption of this new metho because of the ease and economy of time in correcting papers. For that matter, much of the correcting work can be turned over to a clerk without danger of decreased accuracy of grading. The traditional examination procedure involved little attention to the preparation of the actual questions,

ut it did involve a great deal of attention, time, effort, and xpertness in evaluating the bewildering variety of answers. The new-type examination reverses the process. *Expertness n preparing the examination is substituted for expertness in orrecting the papers.* Furthermore, the efforts formerly xpended in correcting the papers were lost, whereas the fforts now expended in preparing new-type questions can e preserved; that is, a comprehensive file of a thousand uestions can be built up for each subject so that future xaminations can be assembled with little effort by simple eference to the file.

The foregoing principles are basic in a consideration of the roblem of measuring pupil achievement in school subjects. Vhile it is true that they seem obvious after slight reflection, et their application in actual examination practice requires nsiderable ingenuity and thought on the part of the teacher ho desires to put them into practice. A careful study of ie material that follows should reduce the difficulties of ccessful application of these principles to a minimum, the etailed description of the different forms of new-type uestions and the rules for formulating and using them being ganized with this specific aim in mind.

IV

Common Forms of New-Type Questions

New-type questions may be divided into two general classes: (1) the *recall* type and (2) the *recognition* type, each having its different varieties as shown in the following outline.

Varieties of New-Type Questions

Recall type (one-word answer originated by examinee)
- Ordinary question . . . . . (*a*)
- Completion form . . . . . (*b*)

Recognition type (choice of given alternative answers)
- Choice between two alternatives (True-False) . . . . . (*c*)
- Choice among three or more alternatives
  - Single choice . . . . . (*d*)
  - Plural choice . . . . . (*e*)
  - Matching . . . . . (*f*)
    - Special form of question (analogies) . . . . . (*g*)

By the recall type is meant the form of question to which the examinee supplies his own answer, as distinguished from the recognition type in which two or more alternative answers are furnished and the examinee has merely to choose the right answer from among the given alternatives.

Questions of the recall type are almost always to be answered by one word. The form of the question may vary, however. The two principal forms are the ordinary question (*a*) and the so-called completion form of question (*b*). Both these forms will be described and illustrated below.

There are various forms of the recognition type of question, depending upon the number of alternative answers from

which the choice is made, the number of choices to be made from the alternative answers, and the form of the question itself. The natural division may be made between questions having but two alternative answers and those having more than two, for the reason that the former kind involves a relatively large element of chance and for that reason usually involves a specialized type of scoring called the "right-minus-wrong" method. Of those questions having two alternative answers, the most important type is the so-called "true-false" type (*c*), which will be explained and illustrated below.

The questions involving a choice among three or more alternatives may be divided into sub-classes according to whether a single choice is made (*d*), whether two or more of the alternative answers are to be chosen (*e*), or whether the answers are indicated by matching answers with questions, questions and answers being given in parallel columns (*f*). A special form of question known as the analogy (*g*) deserves particular mention, although this may be answered in any one of the ways mentioned above. In the majority of cases, however, analogies are answered by a single choice of four or five alternative answers.

In the following pages these types are described more in detail and are illustrated.

(*a*) ONE-WORD-ANSWER RECALL TYPE

This form requires as a correct answer a single word or phrase to be written by the student. Suppose that we desire to test a pupil's knowledge of a barometer. The usual form of question would be, "What is a barometer?" The one-word answer would be, "What is the name of the measuring device which is used to measure air pressure?" Here, only one answer is correct; i.e., barometer.

ADDITIONAL ILLUSTRATIONS OF THE ONE-WORD-ANSWER RECALL TYPE OF QUESTION IN A FEW SELECTED SUBJECTS

Arithmetic. If it takes 3 men 2 days to dig a trench 50 feet long, how long would it take 2 men to dig a similar trench? Answer 3

History. The World War Peace Treaty of 1919 was signed at Versailles ..... (10)

Physiology. The end organs of taste occur in the mucous membrane of the tongue ..... (6)

Physics. What is the name of the law governing the expansion of gases under constant pressure? Answer Boyle's Law

English. What character portrays the relentless Jewish money lender in Shakespeare's *The Merchant of Venice?* Answer Shylock

Economics. The law of supply and demand applies chiefly to price ..... (5)

Chemistry. The formula for nitric acid is HNO_3

Astronomy. The Great Dipper is in the constellation of Ursa Major

Civics. What is the name applied to the measure giving the people the right to approve or disapprove of legislative acts?

Answer Referendum ..... (10)

French. What does "fenêtre" mean in English? Answer window ..... (6)

The practice of indicating the number of letters in the correct answer is recommended unless such indication itself gives too obvious a clue to the right answer, in which case simply a blank space or a line is left.[1] For example, "What is the name of the measuring device which is used to measure air pressure? Answer Barometer ..... (9)." This form of presentation enables the pupil to check immediately the

[1] The members of a large class in elementary psychology (226 students), having been exposed to both the dot technique and the blank spaces, were asked to vote their preference. The result showed 55 per cent preferring the dots, 34 per cent not preferring the dots, and 11 per cent indicating no preference.

correctness of the answer that occurs to him, at least so far as the number of letters in the correct answer is concerned. Furthermore, it discourages guessing and encourages definiteness of expression. As a result, the variety of possible answers is reduced to such an extent that keys of acceptable answers can be prepared by the teacher, so that assistants (some of whom may not even know the subject matter) can readily and accurately score the examination papers.

The dot technique does not reduce the variety of answers to one, and indeed does not necessarily reduce the number of correct answers to one only, for, in spite of great care in framing the question, the teacher will discover that some pupils will be able occasionally to produce an equally good answer which is longer or shorter than the one intended. The writer advocates that such equally good answers be given full credit, even though they do not have the required number of letters.

The example given above illustrates the use of the one-word-answer recall type of question in testing knowledge of definitions. Instead of asking a pupil to give a definition of something, he is given the definition itself and asked to indicate what it defines. This form of question is equally well adapted to test knowledge of dates, events, authors, characters, experiments, etc. These illustrations seem limited to matters of information and memory; yet the use of the one-word-answer question to test ability to apply principles, to make comparisons in terms of similarities and differences, or to give reasons is not impossible. It is true that such questions are more difficult to prepare, but the difficulty is not necessarily due to the one-word-answer form of question. Such questions usually are more difficult to prepare in general, regardless of the form of question. An illustration of the one-word-answer question involving more than mere information and memory is as follows:

Testing knowledge of a principle. New-type question: "What volume at 0° C. would 10 liters of oxygen gas at 30 C°. occupy? Answer _____ degrees C." The definite answer required is 9.0099 and cannot be given unless the pupil knows not only the general principle involved in Charles' Law but also how to apply that principle to a concrete problem. It goes without saying that failing this question would not necessarily indicate lack of knowledge of the principle or law itself; so it would be desirable to precede this applicational question by a question testing knowledge of the principle itself.

Perhaps a further word is necessary concerning the assumed opposition between memory or information and reasoning. A quotation nicely illustrates the point: "There is not as much opposition between 'information' and 'reasoning' as some teachers would have us believe. Facts do not exist in the mind in isolation. We remember by thinking, and we think by remembering facts. Those who declaim against tests of 'mere facts acquired and remembered' should realize that we cannot think without facts. We could not neglect the measurement of facts without neglecting the measurement of the very fabric of thinking. When we consider that facts are not only a legitimate and undoubted aspect of thinking, and also that they can be acquired, retained, and reproduced only by thinking, only by organizing material in a logical and systematic manner, there can be no doubt of the value of the pure information test. In fact, the measurement of reasoning and organizing ability in a field of knowledge would be very defective and incomplete without a measurement of the breadth of information in that field. . . . Every experimental study thus far made and reported has shown a very high relationship between measurements of information in a field and intelligence or ability to think in the material of that field." [1]

[1] Ben D. Wood, *Measurement in Higher Education*, pages 162–163.

General rule for preparing ordinary questions of the one-word-answer recall type. Write out what you consider to be an adequate and concise answer to the question you would ordinarily ask and then use this answer in framing a question which calls for a single-word answer (or, at most, two short words) — that is, a key word which is vital to an adequate understanding of the question. For example, you desire to test a pupil's knowledge of the term "value" as used in elementary economics. You would ordinarily state the question, "What is value?" and would accept this concise answer as adequate, "Value is usually used to denote the rate at which any two commodities are freely exchanged." You can now use this answer in framing your question as follows: "What denotes the rate at which any two commodities are freely exchanged? Answer *value* ..... (5)."

It is sometimes desired to have the pupils give an answer requiring two or more parts, in which case it is often possible to state the question in the form described above and to provide spaces for giving the two or more single-word or single-phrase answers. Questions involving the listing of a number of terms, or of a number of characteristics of an event or process, etc., are easily arranged in this form.

(*b*) COMPLETION FORM OF RECALL TYPE

This form of question originated as an intelligence test many years ago, when the German psychologist Ebbinghaus first proposed its use as a psychological test method. It consists of the preparation of a statement with certain words omitted, the requirement being to supply the missing words so as to make the statement sensible. For example, "Many abnormal symptoms are only *exaggerated* ..... (11) forms of mental processes found in normal persons." To complete this satisfactorily requires a knowledge of the

nature of abnormal symptoms as studied in abnormal psychology and also of the fact that a given mental process varies quantitatively from a minimum degree to a maximum degree. The student who fully understands the import of these facts would have little difficulty in supplying the missing word "exaggerated" to complete the sentence. Of course, a student who had memorized this particular statement in advance, even though he had little knowledge of the implications, would be able to supply the missing word on the basis of sheer rote memory. This possibility makes it unwise for a teacher to burden an examination with incomplete sentences and incomplete statements taken verbatim from textbook or lectures. To avoid "parrot-like" memory answers it is necessary to paraphrase or re-phrase important principles, methods, or facts when putting them into completion form.

ADDITIONAL ILLUSTRATIONS OF THE COMPLETION TEST METHOD

Geometry. The part of a circle included between two radii ..... (5) and an arc ... (3) is called a sector.

Physiology. Harvey discovered the circulation ..... (11) of the blood.

Geography. The French colonists settled the interior of America by using water transportation via the Great ..... (5) Lakes ..... (5).

Physics. Boyle's Law states that when a gas is subjected to compression and kept at a constant temperature the volume ..... (6) is inversely ..... (9) proportional to the pressure.

English. Dickens's *Tale of Two* Cities ..... (6) tells of the French ..... (6) Revolution.

History. "54–40 or fight ....." (5) was a slogan used during the settlement of the Oregon ..... (6) boundary dispute.

Algebra. $(a + b)^2$ is equal to $\frac{a^2}{\quad} + \frac{2ab}{\quad} + \frac{b^2}{\quad}$.

French. Word-order in French sentences usually requires that the subject .... (7) stand first, followed by the predicate ..... (9).

General rule for preparing the completion form of recall-type questions. Here, again, make a list of the questions you would ordinarily use in preparing an examination; then write out adequate concise answers, using these answers as the basis for your new-type completion question. In using these concise answers, eliminate the words that seem to be the most important, making certain at the same time that the remaining words and phrases are sufficiently complete to suggest the proper missing words to one who is thoroughly familiar with the principle, fact, or method stated. This procedure can be illustrated by utilizing the question and answer about value mentioned above. The completion form of handling the answer describing the meaning of the term "value" would be something like this: "Value is usually used to denote the rate .... (4) at which any two commodities ..... (11) are freely exchanged ..... (9)." The assumption here is that a good pupil could have given the essentials of the meaning of value in some such language and hence, if he knows the whole, will be able to complete a partial or incomplete statement by supplying the missing parts.

(*c*) TRUE-FALSE FORM OF RECOGNITION TYPE

Discussion of this form is deferred, since its significance is better understood after the discussion of the form involving a choice among three or more alternatives.

(*d*) SINGLE-CHOICE FORM OF RECOGNITION TYPE

This type of question presents a statement, together with several (usually four or five) alternative answers, only one of

which is correct. The pupil is thus required to exercise his judgment as to the right answer by underlining or otherwise indicating which of the several alternative answers he considers to be correct. It is a test of his ability to recognize the right answer and to make a choice among several alternative answers. For example, "What denotes the rate at which any two commodities are freely exchanged? (labor cost; demand; value; supply)." The pupil is given credit if he underlines "value" and is given no credit if he underlines one of the other three words, if he fails to underline any of the four words, or if he underlines any two. The assumption here is that the pupil who knows what value is will readily recognize the statement as a description of value and will answer accordingly. On the other hand, it is assumed that the pupil who does not know what value really is will find the alternative answers equally or more plausible and hence will underline the wrong one. The chance of getting the correct answer on the basis of guessing is only one in four mathematically; hence effort and ingenuity should be utilized to reduce further this chance of guessing correctly. This can be done by including among the alternative answers those which seem plausible and yet are unequivocally wrong. The alternative, wrong answers which seem plausible to the person who is uninformed must really be wrong answers. This point is emphasized because it is easy for partially correct or possibly correct answers to creep into these alternatives in the effort to make them plausible. When this occurs, the value of the question is destroyed, for it gives rise to disagreement and argument as to what really constitutes the correct answer.[1]

[1] Some teachers deliberately attempt to include among the alternatives terms which are partly correct but which are not so good as the intended correct term. They feel that this scheme is a more delicate measure of judgment or reasoning. The writer has no quarrel with the attitude, but he feels that a teacher should have had a great deal of practice and experience in preparing single-choice questions

Instead of having the pupil merely underline the correct word or phrase, it is much better to number the answers and have the pupil place the number in parentheses at the right-hand margin of the page, in order that the responses may all come in a single column to which a key may be applied much more easily than to underlinings scattered here and there over the page. It is also desirable to have the pupil indicate the correct answer by underlining it, to serve as a reference in case any of the numbers he places in the margin prove to be illegible and as an aid to the pupil in reviewing his work. This device was invented by Arthur S. Otis and was first used in employment tests prepared for Cheney Brothers' Silk Mills in 1919. Otis later used it in his General Intelligence Examination and then again in his Self-Administering Tests, and it is now coming into quite general use.

ADDITIONAL ILLUSTRATIONS OF THE SINGLE-CHOICE METHOD

English. Mark Twain wrote ((1) Pilgrim's Progress; (2) Ivanhoe; (3) The Call of the Wild; (4) The Gold Bug; (5) Innocents Abroad) . . . . . (5)

Astronomy. Planets move around the sun in orbits that are: ((1) circular; (2) elliptical; (3) hyperbolic; (4) cylindrical) . . (2)

Physical Geography. Among the features due to stream erosion are: ((1) mountains; (2) plains; (3) canyons; (4) glaciers) . (3)

Biology. Man and other mammals show the greatest resemblance as: ((1) adults; (2) infants; (3) youths; (4) embryos; (5) seniles) . . . . . (4)

before adopting this technique. Until such experience is acquired, it would seem better for the teacher to follow the suggestions given above. However, if it is desirable to include partially correct answers for any reason, the teacher should frame the question to read, "Underline the word or phrase that makes the *best* or *truest* answer."

Chemistry. A gas which supports combustion is: ((1) nitrogen; (2) carbon dioxide; (3) hydrogen; (4) oxygen) 4

History. The date of Magna Charta is: ((1) 1492; (2) 1066; (3) 1620; (4) 1215; (5) 1776) 4

Economics. The Malthusian theory deals with the: ((1) fiscal problem; (2) specie payment; (3) population problem; (4) divorce evil) 3

French. "Rentraire" means: ((1) to read; (2) to play; (3) to sleep; (4) to return; (5) to darn) 5

Algebra. If $x - 7 = 12$, then $x =$((*a*) + 5; (*b*) − 5; (*c*) + 19; (*d*) − 19; (*e*) − 7; (*f*) − 12) . . . . . e

Arithmetic. If you can buy pencils at the rate of two for five cents, then you can buy ((*a*) 20; (*b*) 10; (*c*) 100; (*d*) 25; (*e*) 250) pencils for fifty cents . . . . . a

General rule for preparing the single-choice form of recognition-type question. The general rule for preparing this form of question differs in no essential from that noted for the various forms of the recall type as described above. The technique differs only in the ingenuity that must be exercised in making the wrong alternatives as seductive as possible. Special care must be taken to make these plausible and therefore seductive alternatives wrong beyond question.

(*e*) TRUE-FALSE FORM OF QUESTION

The true-false form of question is the most common of the various forms of the recognition type having but two alternative answers. Among the other forms is the one involving a choice between "yes" and "no," as, for example, in Test 6 of the Terman Group Test of Mental Ability, of which the

first item is, "Are cartoons made by cameras? Yes No." Another form is that used in Test 3 of the same Terman test, in which the first item is,

alert — sluggish . . . . same — opposite

The true-false form of question has been used so commonly in the new-type examinations, that many have made the error of making no distinction between the new-type examination and the true-false question. As is evident from the foregoing description of different kinds of new-type questions, the true-false kind must be considered as only one of many forms of new-type questions. As the name indicates, the true-false question consists of a statement which the student must judge true or false. There is a variety of ways of having the pupil express that judgment; sometimes the pupil is required to place a plus sign before the statements he considers true and a minus sign or a zero before those he considers false; or the letters "T" and "F" or the words "True False" may appear before each of the statements, the pupil being required to underline or to encircle "T" or "F" or "True" or "False" for each. If the pupil is required to place a sign before the statement, then the method requiring the placement of a zero before the wrong statement is especially recommended, because the discrimination between the plus signs and the zeros is more easily made; hence the scoring is speeded up without any decrease in accuracy. The encircling of the letters "T" or "F" or the words "True" or "False" before the statements is perhaps the better method. And it is better to have the pupil encircle or underline these than it would be to require him to write the letters or words in a column, partly because such letters are easily confused in scoring and partly because it is easier to score a paper when the answers zigzag than it is when they are in a

single column, for they take on the nature of a profile which the scorer can soon partially memorize.

ILLUSTRATIONS OF TRUE-FALSE QUESTIONS

Civics.	T. (F.)	The U.S. Supreme Court Justices are elected by the people.
	(T.) F.	All appropriation bills in Congress must originate in the House of Representatives.
Geography.	(T.) F.	Detroit, Michigan, is one of the most important automobile-manufacturing centers.
	T. (F.)	The chief crop in Ohio is tobacco.
Physiology.	(T.) F.	The normal pulse rate is about 70.
	T. (F.)	Pepsin is secreted by the thyroid gland.
Geometry.	T. (F.)	The intersection of two planes forms a point.
	(T.) F.	Any chord passing through the center of a circle is called a diameter.
French.	T. (F.)	All nouns in French are either masculine, feminine, or neuter.
	(T.) F.	"Le" is a definite article.

In preparing these statements, some of which are true and some of which are false, the opportunity for exercising considerable ingenuity presents itself. The false statement, as a general rule, must not be so obviously false that a relatively uninformed person would recognize its falsity. For example, in preparing a statement about "false beliefs," advantage is taken of the known fact that pupils tend to confuse the terms "delusion" and "hallucination." The better pupils acquire a clear-cut knowledge of the difference between the two terms, whereas the poorer pupils frequently fail to see the distinction. With this in mind, the statement "False beliefs are called hallucinations" is placed before the pupil. The better pupils know that an hallucination is not a false belief but is a false sensory experience, and they also know that a delusion is a false belief; so they are able to

utilize this knowledge in judging that the statement before them is false. The poorer pupils, lacking this clear distinction, are inclined to mark it true because the word "false" which characterizes both delusions and hallucinations is here used with the word "hallucinations."

In spite of such ingenuity as may be exercised to mislead the less-informed pupil, there may be still a large element of guessing. That is, a completely ignorant person can mark 50 per cent of the statements correctly by chance. Therefore, if true-false statements are to be used, it is necessary to utilize a very large number of them and to adopt a scoring method that will counteract this guessing factor. Discussion of this point will be presented in a later section of this manual.

(*f*) PLURAL-CHOICE FORM OF RECOGNITION TYPE

This type of question is merely an extension of the single-choice recognition form. It has been devised for use where the questions require two or more items in the answer. These two or more correct answers are given along with a number of unequivocally wrong but seductive alternatives. Two examples will suffice to show that this form of question is very similar to the single-choice form of the recognition type.

In the following list underline each word that is a chemical compound: (air; water; nitrogen; ammonia; sugar; argon; bismuth).

The second, fourth, and fifth words in this example are correct and should be underlined.

In the following list underline each city that is the capital of a state: (Pittsburgh, Pa.; Topeka, Kansas; Albany, New York; Cleveland, Ohio; Austin, Texas; San Francisco, Cal.; Minneapolis, Minn.; Frankfort, Ky.; New Orleans, La.; Providence, R. I.).

The second, third, fifth, and tenth cities in the list should be underlined.

(*g*) PAIRING OR MATCHING TERMS IN PARALLEL COLUMNS

This form has been used only to a slight extent in new-type examinations but is mentioned here because it is admirably adapted to testing certain kinds of knowledge. For instance, suppose that in a history course you desire to test knowledge of the chief characteristic of each of a dozen or more administrations and you wish to do this without having the pupil write lengthy answers. In this case you would list in one column in chronological or any other order (logical or haphazard) the various administrations, numbering each, and then in a parallel column you would present, in random order, a list of chief characteristics. The pupil would then be required to match or pair each characteristic with its appropriate administration, indicating this in each case by placing the appropriate administration number before or after its characteristic. The following is an example:

Presidents	*Chief Events*
1. Thomas Jefferson	3 Rise of Spoils System
2. James Madison	5 Dred Scott Decision
3. Andrew Jackson	7 Fourteenth Amendment
4. James K. Polk	12 Building Panama Canal
5. James Buchanan	6 The Civil War
6. Abraham Lincoln	10 Panic of 1893
7. Andrew Johnson	1 Louisiana Purchase
8. U. S. Grant	9 Resumption of Specie Payment
9. R. B. Hayes	11 The War with Spain
10. Grover Cleveland	8 The Credit Mobilier
11. William McKinley	4 The Mexican War
12. Theodore Roosevelt	2 The War of 1812

(*h*) THE ANALOGY FORM OF QUESTION

This form is given special mention because of its difference from the ordinary question. It is especially well adapted to test a pupil's ability to see fundamental relationships between various items studied in a course. The question is based upon an analogy such as "Day is to night as white is to black." In giving the question, however, one of the four terms (usually the fourth) is omitted. The pupil may be required to supply the missing term, making it a recall-type question; but usually four or five alternative answers are given, only one of which is correct, and the pupil is required to choose the correct answer. The analogy form of question is, therefore, of the recognition type. The following are samples of this form:

Automobile is to carriage as motorcycle is to (walking, horse, buggy, bicycle, train).

Circle is to square as sphere is to (circumference, cube, round, corners, ball).

Gastric juice is to the stomach as (saliva, adrenin, tears, bile) is to the lachrymal glands.

Note that grammar is to be sacrificed in some cases, as, for instance, in the third example just given.

This form of question loses all meaning unless the first two terms bear the same relation to each other as the third and fourth terms bear. The precautions noted in preceding sections for the proper preparation of the one-word-answer question and for the recognition form of question also hold in the analogies form.

V

Advantages and Disadvantages of Each Form of Question

The following attempt to list the advantages and disadvantages of each form of question, while not exhaustive, is sufficiently complete to give an inkling of some of their less obvious characteristics.[1]

ADVANTAGES OF THE RECALL TYPE

Success in answering this form of question is dependent upon the capacity to recall and apply principles, methods, facts, etc., which have been thoroughly learned. For this reason it is one of the best of the new-type questions to test a pupil's thoroughness of learning and the degree to which he has organized the theoretical and factual materials in the course.

There is little danger that this type of question will tend to test in a rote manner that which has been learned in recitation drill. It is so difficult to prepare this form of question that teachers will continue to quiz pupils orally with the old "how" type of question. Hence pupils will not be coached specifically in oral quizzes to answer examination questions of this type.

This form of question forces the pupil to be brief, concise, definite, and specific in thinking out and phrasing his answers. Such questions lead the pupil to give greater attention in study and preparation to the organization of subject matter and to the correct *understanding* of the details as a basis for

[1] No specific rules can be laid down to govern the question as to how large a class should be before it becomes "practicable" or "economical" to use objective examinations. This manual has been prepared primarily with the needs of large classes in mind. However, some teachers use objective examinations with classes as small as ten or twelve. The writer's bias in favor of this form of examination leads him to advocate their use whenever the accurate measurement of pupil achievement is considered to be really important. Hence he would advocate their use with small classes as well as with large classes.

accurate generalization. Thus halfway measures in learning tend to be discouraged and pupils are led to depend less on vague ideas, general impressions, and empty generalities.

Brevity in this form of question coupled with brevity in the required answers permits the use of a relatively large number of questions without crowding the pupil for time. This results in a very decided advantage as far as permitting an adequate sampling of every phase of the course is concerned. By utilizing this type of question in large numbers the teacher is able to sample or test each pupil's knowledge of every phase of the course content, thus necessitating, on the pupil's part, consistent application to and study of each textbook assignment, each reference reading, and each class discussion.

Ease in scoring must also be listed as a distinct advantage possessed by the one-word-answer question. By controlling the correct answers and so framing the question as to permit of but one correct answer, it is possible to make a key for scoring the questions so that the routine scoring can be turned over to a clerk who need know little or nothing concerning the course content itself. Such a key gives to the examination a series of definite units, each to be credited with one or two points when answered correctly. Because of the number of questions and the adequacy of the sampling of subject matter, it is possible to disregard the usually vexatious problem of weighting questions.

DISADVANTAGES OF THE RECALL TYPE

The chief disadvantage lies in the difficulty of preparing this type of question. Teachers find it very difficult to abandon suddenly their habits of phrasing traditional questions. Not only is it a matter of breaking long-standing question-framing habits, but it is also a question of the time and effort required to write out adequate yet concise answers as a basis for framing this type of question. This disadvan-

tage becomes less serious as one develops skill in framing such questions. Such skill develops, of course, in proportion to the amount of practice one has in preparing such questions.

Another disadvantage may be said to exist in a somewhat natural tendency to base many of the questions on minor and relatively unimportant details of the course and to stress matters of rote memory rather than to stress items involving meanings, principles, applications, and the like. Here, again, the disadvantage tends to disappear as the teacher acquires skill in selecting and preparing thought-provoking questions bearing on the more fundamental and important phases of the subject matter.

One drawback to the single-word-answer question is its lack of inclusiveness. In testing knowledge of some involved process, it is impossible to select one key question that will infallibly indicate the extent to which any given pupil knows the whole process. To overcome this disadvantage requires either the framing of a series of key questions each covering a phase or part of the total process (and there is certainly no objection to doing this) or adopting the expedient of framing the question so as to call for a multiple answer. In resorting to this expedient it is merely necessary to follow the directions governing the preparation of the one-word-answer question involving a multiple answer. Since this variant of the one-word-answer method is practically the same as the one-word-answer type of question in principle and method of construction, it will not be discussed further here.

THE COMPLETION TYPE

The disadvantages of the completion type of question are somewhat less than for the ordinary recall question, for it is relatively easy to cover a whole complicated process by describing it and leaving blanks at critical points in the description. It is also somewhat easier to prepare. Good

judgment in determining the parts to be left blank is necessary, and, since no rules or principles are available to serve as a certain guide, it is best to submit a proposed completion test to several teachers for independent checking so as to secure the benefits of a combined judgment in preparing the final test for actual use with a class.

It is probably true that a pupil's success in handling the completion form of examination is dependent to a considerable extent upon his general intelligence. However, a person rating very high in standard intelligence tests but lacking in knowledge of a given subject would be certain to do poorly in a completion examination in that subject. On the other hand, of two pupils equally well prepared in the subject, the one who rates higher in standard intelligence tests would probably make the better mark. Some would classify this fact of dependence on general intelligence as an advantage, whereas others would consider it a disadvantage. If we define the function of an examination to be the reliable measurement of "knowledge of, *and ability to think in*, the material of the course," [1] then we should be inclined to classify this dependence of the completion question upon general intelligence as a decided advantage.

ADVANTAGES OF THE RECOGNITION FORM

The advantages of this form of question are similar to those mentioned for the recall-type question. It is an easier type of question to answer, for we are able to recognize much more on the basis of previous experience (study, hearing lectures, etc.) than we can actually recall. We are able to recognize the meaning of many more words than we are able to recall and use in sentences of our own construction. It is just this fact of being able to recognize more than we can

[1] Ben D. Wood, *Measurement in Higher Education*, page 153.

recall that makes it wise to use this form of question as a supplement to the recall or single-word-answer form; it gives a measure of the pupil's "acquaintance with" the many phases of the subject matter, whereas the recall type of question is more likely to give a better measure of the pupil's "knowledge about" the subject matter. This comparison or contrast is not to be taken too literally, for the recognition form is not limited to measuring simple "acquaintance with" the subject matter. But it is true that the latter is generally easier to answer and hence a pupil may be expected to answer in a given time a much larger number of recognition questions than of recall questions.

This type is somewhat easier to prepare and is much more easily scored, for you need only compare the underlinings with the correctly marked key, or place a correctly numbered key alongside the numbers written in the column at the right-hand side of the page, without even reading any of the words on the examination page.

There is a decided advantage in the use of several alternative answers as compared with the use of only two, as in the case of the true-false type, because the chance of guessing the correct answer is less (i.e., only one out of four instead of one out of two), depending of course upon the number of alternative answers given. When this type of question was first used, it was the common practice to use four alternatives, one being the correct answer, but of late there is a strong tendency to use five alternatives, one of them being correct. Sometimes even as many as seven alternatives are used.

DISADVANTAGES OF THE RECOGNITION FORM

One disadvantage is found in the tendency to make the alternative incorrect answers partially or possibly correct.

This results in ambiguities and disagreements concerning the correct answer. Such ambiguities are very likely to creep in because of the teacher's efforts to make the incorrect alternatives as plausible and seductive as possible. The only remedy for preventing this unfortunate error is to submit the questions to another equally competent teacher for trial, his job being to take the test. A check-up on his answers in comparison with the supposedly correct answers designed by the one preparing the examination will reveal those questions about which there is disagreement due to ambiguities either in the statement or in the possible answers. *The practice of making such an independent check of the answers cannot be too strongly recommended. This method is followed by those who have had most experience in preparing and giving such examinations.*

ADVANTAGES OF THE TRUE-FALSE FORM

On first sight the answering of true-false statements seems relatively simple. But the mental processes involved in making the judgments concerning the truth or falsity of those statements are by no means so simple. They demand the application of learned facts and principles to new situations and may involve thinking or reasoning of a high order. The extent to which such a question evokes involved reasoning is dependent, of course, on the ingenuity of the teacher in preparing thought-provoking statements. Such ingenuity apparently is itself dependent in large part on experience and practice in preparing true-false statements.

Another advantage inherent in this type of question lies in the speed with which a pupil may answer these questions; hence a relatively large number may be given in a short period of time with a resultant thorough sampling of every phase of the subject matter. From eighty to one hundred and fifty may be given in an hour examination. The time

for checking or underlining the "True" or "False" printed before each question is of course negligible, and the time required for simply reading the questions, while not negligible, is very short. This means that the bulk of the hour may be given over to genuine thinking, the results of such thinking being indicated without the use of writing.

It is obvious that these true-false questions are very easily scored or marked by the use of a stencil or correctly marked key. The correctness or incorrectness of the underlining or encircling may be noted and indicated opposite each question without reading the questions at all, comparison with the stencil or correctly marked copy alone being necessary.

DISADVANTAGES OF THE TRUE-FALSE FORM

The chief difficulty with this form of question, as with any of the forms calling for a choice between only two answers such as "yes-no" questions, arises because of the "fifty-fifty" chance of guessing the correct answer, though it is true that the seductiveness or the obviousness of the statement affects this guessing element so markedly that we cannot assume the guessing factor to be present to this extent. Nevertheless, there is sufficient evidence on record to indicate that this guessing element is present to such a degree as to render a surprisingly large proportion of questions valueless. This being true, we are forced to utilize a very large number of questions in order that the group of true-false questions may differentiate between the good pupils and the poor pupils. This simply means that the value of a true-false test is to be found in the relatively small number of questions which really differentiate between the good and poor pupils while the rest of the questions, even though they may not be detrimental, constitute so much "excess baggage." Most workers have recognized the

necessity for overcoming this disadvantage by using a very large number of true-false statements whenever this form of question is used at all. The recommended number is seventy-five as a minimum, with one hundred as the more desirable number. Wood, in a recent study of the use of this type of question in law examinations, found this recommended number to give results not sufficiently reliable for his purpose and hence recommended the use in such examinations of two hundred questions as a minimum. These recommendations make this form of test somewhat unwieldy, especially when allowance is made for the fact that other forms of new-type questions should be employed as well.

Since the use of a large number of true-false statements is not sufficient to counteract the effects of guessing, a method of scoring commonly known as the right-minus-wrong method is frequently used. This method is also used for the other forms of recognition test having but two alternative answers, such as the "yes-no," "same-opposite" forms, etc.

The way in which the scoring method counteracts the effect of guessing is as follows: Suppose that 100 statements are included in a true-false test and that a pupil has positive knowledge of 50 of them and no knowledge whatever of the other 50. At first thought the reader might assume that his score in the test would be 50 (the number of statements about which the pupil had positive knowledge); but it should be remembered that since there are only two ways of answering any question, even if the pupil has no knowledge of it whatever, he is as likely to guess right as to guess wrong, so that in thus guessing at the 50 about which he has no knowledge whatever, he will most probably guess 25 of them right. He may guess somewhat more or somewhat less than exactly half of them right, but he is more likely to guess 25 right than any other number. This means that his score instead of being 50 is most probably 50 plus 25, or 75. Now,

since the number that he got right by pure guess is most probably equal to the number wrong, if we subtract the number wrong from the total number right, this will have the same effect as if we subtracted from the total number right those which were got right by pure guessing and leave a score equal to the number that were got right because of the student's having positive knowledge. Thus, in the case we have cited, subtracting 25, the number wrong, from 75, the number right, would give 50, which is the number of questions about which the pupil was supposed to have positive knowledge. In using this scoring method, questions omitted by the pupil are not counted as wrong, for it is obvious that if the pupil does not guess at the answer to a question but leaves it blank, he could not get a right answer.

The above example is stated as if the number of wrong answers is always equal to the number of answers guessed correctly. Obviously this is not true, the facts being of the nature of probability. The scoring formula, based as it is on probability, represents the best estimate under the circumstances, this estimate being close to the facts in each individual instance in proportion to the square root of the number of questions. Hence the formula applied to 1000 questions would be closer to the true score for each person than it would be if applied to 100 questions. But it would not be ten times as close, being, in this instance, only three times as close. Such reasoning is based on the theory of pure chance and does not hold strictly in true-false examinations because of the seductiveness or the obviousness of the questions and because the pupils are not completely ignorant of the facts covered by the examination. The writer's opinion is that, in a 100-question examination, the student's score would not be in error on the average by more than five or six points. The error is far less than a layman would ordinarily anticipate, at least if that layman has the preva-

lent exaggerated notion of the accuracy of traditional examination methods.

The score by the "right-minus-wrong" method may be found more quickly in some cases by subtracting twice the number wrong from the number of items attempted. For example, suppose that in a 100-item true-false test a pupil attempts 90 items, having omitted or left blank 10 items, and gets 15 wrong. His score would be 90 attempts minus 30 (twice the number wrong), which would be 60. Stated the other and longer way around, this pupil omitted 10 items, answered 75 items correctly, and answered 15 items incorrectly. The 75 rights minus the 15 wrongs yields the same score; i.e., 60.

The true-false type of question suffers from still another disadvantage similar to that mentioned for the single-choice type; namely, the tendency to include statements which are partly true and partly false. The writer knows of at least one instance where a teacher made a dismal failure in using this type of question simply because his examination contained many ambiguous statements. Some statements thought to be true by many pupils were marked wrong by the teacher, and some statements considered false by many pupils were considered to be true by the teacher. When the papers were handed back, some of the pupils were not convinced that their papers had been correctly graded. They consulted with another teacher, asking his opinion concerning the truth or falsity of the disputed questions. Unfortunately (for the examining teacher) this consultant disagreed with the original teacher's answers, with a resultant dissatisfaction that was far from wholesome. Here, again, prevention of ambiguities is accomplished by submitting the prepared list of true-false statements to two or more teachers for independent judgments concerning each item. In this way ambiguities may be detected and the items either modified or dropped from further consideration.

THE PLURAL-CHOICE FORM OF RECOGNITION QUESTION

The chief advantage of the plural choice of answers is that the chance of guessing the right answer to the item is reduced. For example, there is one chance in five of a pupil guessing the right answer in the case of a single choice among five alternatives, but if two answers out of the five must be chosen, there is then only one chance in ten of the pupil getting the item right by guessing both of the answers.

Two answers, of course, take slightly longer to score than one. The use of two choices also prevents the convenient use of numbering the answers and having the numbers placed in a column at the right-hand margin of the page.

Apart from these, the advantages and disadvantages of the plural-choice form of question are approximately the same as those for the single-choice form.

ADVANTAGES OF THE ANALOGY FORM

The peculiar advantage possessed by this form of question lies in the possibility of testing a pupil's knowledge of relationships existing between various principles, theories, facts, etc., in the course. Thorough mastery of the isolated parts of the course is usually essential for the perception of relationships. The pupil who is able to think of the various aspects of the course in terms of their relationships, their similarities, and their differences is certainly superior to one unable to do so. This form of question, therefore, would seem to be almost ideal for testing "knowledge of, and *ability to think in*, the material of the course."

This form of question, in addition, possesses the advantages already noted for the single-word-answer and the single-choice form of questions, these being utilized in the formation of the analogy type of question.

DISADVANTAGE OF THE ANALOGY FORM

The chief drawback to this form of questions is found in the difficulty of preparing them. At least, teachers in psychology have found it a very difficult task, but their persistence has been rewarded by the final production of a fairly large number of analogy questions covering the elementary course. Even a small number of those questions are worth including in an objective examination because our analyses, as far as they go, indicate that a surprisingly large proportion of such questions actually differentiate between the superior and the inferior students.

ADVANTAGES AND DISADVANTAGES OF PAIRING OR MATCHING TERMS IN PARALLEL COLUMNS

This form possesses the same advantages as the recall form involving several answers, since it is similar in make-up. This form, however, tests recognition or "acquaintance with," whereas the other form tests ability to recall or "knowledge about." No writing is involved in answering this form; hence its scoring is somewhat more objective. This form is not desirable when the list of terms is few in number, because of the possibility of guessing. This disadvantage disappears if the number of pairs in the question is as large as twelve or fifteen. At first sight it might appear to be a disadvantage to have as many as twelve pairs in the question, because of the possibility of eye strain and the waste of time in looking all the way up and down a long column. This disadvantage is more fancied than real. The writer included in a recent examination a matching question of twenty-seven pairs, none of the pupils complaining of discomfort or waste of time.

VI

Directions for Preparing and Using Objective Examinations

The following directions are based largely on the writer's experience in preparing objective examinations for his own courses,[1] in aiding the staff to prepare objective examinations for the eight or nine hundred students in elementary psychology, and in consulting with other faculty members and teachers who sought aid in preparing the new-type examinations. In addition, these directions are based on general principles of examining derived by those who have experimented in the development of standard intelligence tests, trade tests, and educational measuring scales. As far as possible the reasons for the various directions will be stated, but even so the writer fears that they may seem dogmatic.

1. In assembling questions, care should be taken to cover every phase of the course.

One of the advantages of new-type questions lies in their brevity and hence the possibility of covering every phase of the course. It is essential to sample widely the contents of the course to be covered by the examination in order that each pupil's rating or grade will represent his mastery of the

[1] The writer has been using objective examinations since 1915, when, as an assistant in the Department of Psychology at Ohio State University, he learned from Dr. A. P. Weiss the technique involved. Reference is made in the bibliography to an article published by Dr. Weiss in 1911, describing the use of the "completion test" method of examining students in introductory psychology at the University of Missouri. One cannot discover the exact origin of such methods, for they were probably developed in one form or another in many places. The present widespread interest shown in schools and colleges, of course, is due to the successful use of objective examinations at Columbia University since 1919, these being made possible by the ingenious devices invented by Arthur S. Otis and successfully utilized in the Army Alpha intelligence tests, which were given to 1,700,000 draftees during the recent war.

course as a whole. The principle here is the same as that which governs the work of the geologist in sampling the ore of a mine or of the grain dealer who samples the grain in a carload lot. The sampling must be thorough and representative of every part of the thing sampled.

The best way to insure thorough sampling is to prepare questions for each chapter studied, for each lecture delivered, for each reference reading assigned, or for each experiment performed. When two or more teachers coöperate in preparing an examination, it is economical for them to divide the various assignments among themselves, each preparing questions on a specified part of the course.

2. *In preparing questions, an effort should be made to secure many more questions than will actually be included in the examination.*

If this be done, ample opportunity will be given for selecting from among the suggested questions those which are most thought-provoking, rejecting those which are based on insignificant details or are ambiguous, poorly phrased, or labored.

3. *Ambiguous questions should be avoided.*

The only safe way to accomplish this is to have two and preferably more teachers answer each proposed question independently, and to compare their answers in order to reject those questions concerning which there is disagreement or to revise them so as to bring about agreement as to the correct answer. Those who have had little experience in the preparation and use of the new type of examination are especially cautioned with reference to this point. Much may hinge on meanings read into the terms used in phrasing

a question, and yet these various possibilities may remain hidden until independent answers given by two or more teachers are compared. Too much emphasis cannot be placed upon the necessity for rejecting ambiguous or poorly worded questions by some such method as is here recommended. Each question should be phrased in such words that there would be no disagreement among competent judges as to their meaning. The same, of course, applies to the answers.

4. *In determining the eligibility of any question, the basis of judgment should not be the apparent difficulty of that question.*

In the first place, judgments concerning the difficulty of a question are prone to error. In the second place, easy questions are just as desirable as hard ones. There is a common tendency among teachers to refrain from asking an easy question because they seem to fear that every pupil will pass it. Experience has shown, however, that frequently a question thought to be easy by a teacher is not so easy when the pupils attempt it. Furthermore, an easy question which actually differentiates between the "D" and "F" pupils on the one hand, and the "C" or better pupils on the other, is just as valuable and indeed just as essential as a hard question which differentiates between the superior pupils and the medium pupils. Moreover, an easy question which differentiates between the poorest pupils and the rest is more valuable than a hard or difficult question which fails to differentiate between the superior and the less superior pupils. Certainly mere difficulty of a specific question, then, should not be the basis for its acceptance or rejection.

5. *In assembling the acceptable questions for the examination, an effort should be made to include an equal number of easy, hard, and moderately difficult questions.*

If a 150-question examination contains fifty easy questions, fifty moderately difficult questions, and fifty hard questions, then there is a good chance for the examination to measure the inferior, average, and superior pupils with equal exactness. It is obvious that an examination consisting only of very difficult questions would fail to differentiate between the inferior pupils and the average pupils, because both groups would fail most of the questions and hence would have similar scores. In like manner an examination consisting only of easy questions would fail to differentiate between the average and the superior pupils because both these groups would pass practically all the questions. Similarly, an examination containing only questions of average difficulty would fail to differentiate between the average pupils and the superior because the latter in such an examination would not be pushed to the limit of their capacity. For these reasons it is necessary to include a sampling of questions of all degrees of difficulty in about equal proportions. Differentiation must be secured all along the line: from those pupils who are so inferior that they will fail some of the easy questions, to those who are so superior that they will pass most of the very difficult questions.

6. *In assembling the acceptable questions, it is well to have the first half dozen or so questions so easy that all pupils will pass them.*

This recommendation arises from the principle of the "shock absorber" theory of testing, which holds that the initial test in a series should be so easy that the person tested is able to do it without effort, thus embarking on the real

test series without initial shock. Compliance with this principle prevents the resentment felt by pupils when they are unable to handle the first few questions and avoids or reduces any nervousness due to apprehension and dread which a pupil may have as he begins an examination.[1]

7. *Each of the acceptable questions should be an independent unit in the examination, not depending on any other question for its meaning.*

Teachers sometimes place a series of related questions in an examination, the answer to several being dependent on ability to answer the first one. If that one be missed the pupil automatically misses the others, although his knowledge of the several elements involved may by no means be as limited as his series of failures would seem to indicate. Or, it sometimes happens that one or more of the related questions is unknown to the pupil, yet its correct answer is either given directly or is strongly suggested by others in the series. These considerations make it wise to build the examination out of independent questions each a unit in itself.

8. *Each of the acceptable questions should be short.*

Other things being equal, the shorter the question the better. This recommendation means that complexly worded questions or statements should be avoided. Unusual or

[1] The attitude of pupils toward this type of examination is important and fortunately has usually been favorable. Data, recently obtained from 226 students in elementary psychology, who had already experienced two one-hour objective mid-quarter examinations and one two-hour final objective examination, show 84 per cent definitely expressing a preference for the new-type examination, 13.3 per cent preferring the traditional written examination, and 2.7 per cent expressing no preference. A larger proportion of those making a high grade in these objective examinations prefer them as compared with the proportion of those making a low grade. However, 72 per cent of those getting an "F" grade in the objective examination actually stated a preference for the short-answer examination.

unfamiliar terms and the use of double negatives should likewise be avoided. If these are not avoided, the question is likely to turn out to be a test of language or vocabulary rather than a test of subject-matter knowledge.

9. The examination should include a very large number of questions.

Within the limits permitted by the time allotted to the examination, the larger the number of questions the better. If only a few questions are included, chance or accident is permitted much greater sway than if many questions are utilized. In other words, the examination should give the pupil as many opportunities as possible to show what he knows. Teachers, experimenting with the new-type examination for the first time, are more likely to err by giving too few than too many questions. One teacher, for example, gave a new-type examination consisting of the ridiculously small number of eight questions. He concluded that the new-type examination is a failure for his particular subject. Not one of those eight questions was significant or valid because each was so easy that approximately eighty to ninety per cent of his pupils passed it. The only thing that this teacher proved was that eight new-type questions, each one valueless, when combined into an examination would likewise be valueless. This teacher was surprised when informed that a fair trial would have demanded at least seventy-five or one hundred questions.

10. Questions should be grouped according to form or type of question, the examination itself consisting of as many parts as there are types of questions.

There are three reasons for this direction. In the first place, each type of question needs to be introduced by

specific directions to the pupil in order to avoid any possible misunderstanding as to what the pupil is expected to do. In the second place, it is possible that a pupil's "set" or attitude toward different types of questions varies with each type; hence he is able to work at higher efficiency if his "set" toward single-word-answer questions is permitted full sway on that type of question before he is required to shift to answer completions or some other form of question. In the third place, scoring of the examination is facilitated and probably made more accurate if the different types of questions are assembled in separate parts. If the papers are marked by two or more persons, each can specialize on a given part of the examination, one person marking the single-word-answer questions, another the single-choice questions, another the completions, and so on. Such division of labor in itself speeds up the marking process, greater accuracy probably resulting from marking or scoring one type of question at a time.

11. Within each part of the examination the questions should be arranged according to topical sequence in the course.

Cognizance is given, here, to the importance of the "mental set" developed by the logical arrangement and presentation of the various topics in the course. Since course outlines usually aim to do this in the hope of instilling logical organization of subject matter, it would seem desirable to maintain, as far as possible, the same arrangement in the examination questions. It is true that we do not have experimental evidence of the wisdom of such arrangement compared, for instance, to the arrangement of questions according to difficulty. Nevertheless, in the present state of our ignorance, it would seem wiser to arrange the questions according to topical sequence in the course.

12. The new-type examination as a whole should be preceded by general directions as distinguished from the special directions preceding each part of the examination.

The aim of such general directions is to give the pupil an attitude of deliberateness and thoughtfulness so that he will not become flustered or worried over inability to answer particular isolated questions. The following general directions have been successfully used in examination procedures:

General directions. "The aim of this examination is to give you an opportunity to show what you have learned from your study of United States history (or whatever the subject may be). It is not expected that you will be able to answer every question. Read the directions for each part carefully and be sure you understand what you are to do before you begin."

13. Specific directions should be given in regard to answering items about which a pupil may be uncertain.

Some teachers penalize for wrong answers without so informing the pupils, whereas other teachers exact the same penalties and do inform the pupils. Some teachers do not penalize for wrong answers and so inform their pupils, while others do not penalize and do not inform their pupils of this fact. With this unfortunate diversity of practice it is no wonder that pupils going from a new-type examination given by one teacher to a similar examination given by another teacher are often in doubt as to whether they should guess or not guess on the questions concerning which they may be uncertain. This diversity of practice is almost certain to continue for some time, until gradually a common practice is established through common understanding and agreement. In the meantime each teacher should announce for each examination, and for each part of each examination,

his plan of scoring. The writer advocates the announcement of the plan of scoring but does not recommend that the pupils be instructed to guess. This position is not based on theoretical considerations, but on grounds of policy. If one definitely advises pupils to guess, then pupils, whose "defense mechanisms" work overtime when they feel they have done poorly, will be justified, to a slight degree, in branding the examination as nothing more than a "guessing contest" akin to a crossword puzzle. It does not seem wise to facilitate such criticism, even though it be ill-founded; hence the avoidance of advice to guess.

Some will take exception to this procedure on the following theoretical grounds: It is better that pupils should guess, because a pupil may know the right answer to several questions but not feel entirely sure, and, if conscientious, he may think he is not living up to directions if he puts down an answer which may be right, when he is not absolutely certain about it, and in this way he would be penalized for not writing down the answers to a number of questions to which he might know the right answers. In other words, guessing enables the pupil to get credit for all that he knows, both thoroughly and partially, for if the pupil does not feel sure of the answers to ten questions but partially knows each one, the chances are that he will get six or seven or even eight of them right even though he thinks he is merely guessing, and in this way will get credit for his partial knowledge. Hence the pupil should be instructed to guess. This position has much to commend it theoretically; yet the writer would not advocate instructions to guess because of the "defense mechanisms" mentioned above. Furthermore, it seems worth while pedagogically to develop an attitude of certainty of knowledge on the part of pupils by teachers, so that pupils will come to know what they know, and what is even more important, will know what they do not know.

However, this whole discussion may be much ado about nothing, since recent experimental work by Ruch and by Paterson and Langlie shows that the right-minus-wrong formula contributes nothing of value to the technique of true-false tests; on the contrary, it actually results in lower reliability coefficients than simply scoring by the number right.[1] In view of these results, indicating that such tests should be scored number right, it would seem wise to discourage guessing by definite instructions not to guess.

The following are sample directions that may prove suggestive and helpful:

Directions for single-word-answer questions. "The questions in this part can be answered by a single word. The number in parentheses after a blank indicates the number of letters in the right answer. If you cannot think of the exact word called for but are reasonably certain that a shorter or longer word is equally correct, then write it down. Your score for this part will be based on the number of questions correctly answered, no penalty being exacted for wrong answers."

Directions for completion method. "The following questions can be answered by a single word for each blank. The number in parentheses after a blank indicates the number of letters in the right word. If you cannot think of the exact word called for but are reasonably certain that a shorter or longer word is equally correct, then write it down. Your score for this part will be based on the number of blank spaces correctly filled. No penalty will be exacted for blank spaces incorrectly filled."

Directions for single-choice recognition form of question. "After or in each statement there are four or more words or phrases in parentheses, each preceded by a number. You

[1] G. M. Ruch, "The Improvement of Written Examinations," and Donald G. Paterson and T. A. Langlie, "Empirical Data on Scoring True-False Tests," to be published shortly in the *Journal of Applied Psychology*.

are to underline the *one word* or *phrase* that makes the *truest* or *best* statement and write its number on the line at the right-hand side of the page. Your score will be based on the number of correct words or phrases underlined and entered in the margin. If a wrong word or phrase is underlined as well as the right one in any given statement, no credit will be given for that statement. No penalty will be exacted for those statements incorrectly answered."

Directions for true-false type of question. "The letters F and T precede each statement below. Encircle the letter F if the statement is false. Encircle the letter T if the statement is true. *Do not guess.* If you are unable to decide whether a statement is true or false, leave it alone. You will be penalized for each statement incorrectly marked by deducting one point or credit from the number of correctly marked statements."

Directions for plural-choice recognition form of question. "Underline *one or more* terms in the parentheses in or following each statement to make the *truest* or *best* statement. *Do not guess.* You will be penalized for each incorrect underlining by deducting one point or credit from the number of correctly marked terms."

Directions for the analogies form of question. "In each of the statements below, the first term *is to* (:) the second term *as* (::) the third *is to* (:) one of the four terms in parentheses. EXAMPLE. boy : man :: girl : (human being; youth; woman; adult). The right word in parentheses is woman, therefore it is underlined. Similarly, proceed to find the right term in the parentheses for each statement below and underline it. Your score for this part will be based on the number of correct terms underlined. If a wrong term is underlined as well as the right one in any given statement, no credit will be given for that statement. No penalty will be exacted for those statements incorrectly answered."

Directions for pairing or matching terms in parallel columns. Directions for this form of question will vary according to the material included, because the directions must indicate what things in both columns are to be matched. The following directions are taken from a psychology examination covering in part certain phases of abnormal psychology. The last two sentences of these directions could be used whenever this form of question is used, regardless of the subject matter involved. "The left-hand numbered list below is made up of instances of normal behavior resembling in principle the forms of abnormal behavior listed in the right-hand list. For each item in the left-hand list there is the name of its principle or class in the right-hand list. Find the principle corresponding to each instance and give it the same number as the instance. Your score will be based on the number of terms in the right-hand column correctly numbered. No penalty will be exacted for incorrectly numbered items." If the teacher desires to counteract the guessing element in this form of question (and he should do so whenever the number of terms is rather small), the last two sentences of the above directions should be changed to read as follows: "*Do not guess.* You will be penalized for each incorrectly numbered term by deducting one point or credit from the number of correctly numbered terms."

14. *The arrangement of true-false items should be a random one, and there should be approximately an equal number of true and false items.*

When true-false statements are used, care should be taken to avoid any regular sequence of true and false items. The safest way to insure a random order is to toss a penny, letting heads represent true statements and tails false statements, arranging the items in the examination according to the outcome of the penny tossing. Ordinarily there should be

approximately the same number of both true and false statements in the test, in order that their arrangement may follow a chance order. This direction holds also for the plural-choice recognition form of question.

15. *The placement of the correct answers among the alternative answers in the single-choice and the plural-choice questions should follow a chance order.*

This means that the correct answer should not always be found in the same position, the order being determined by chance, so that the correct answer sometimes appears first, sometimes second, and so on. To insure a purely chance arrangement for the single-choice questions, one can place in a hat a large number of slips of paper with an equal number of ones, twos, threes, fours, and fives written on them, and then draw out the slips of paper one at a time, the number drawn each time indicating the place among the alternatives of the correct answer in each successive question. To insure a chance arrangement for the plural-choice questions, one can prepare a large number of slips of paper, half containing the word "right" and half containing the word "wrong," and then draw out one slip of paper at a time and arrange the right and wrong answers in the parentheses accordingly. The drawings should be repeated for each successive question, to avoid any definite order from question to question.

16. *The use of a red or a blue pencil in scoring the examination papers and a uniform method of marking correct, incorrect, and omitted answers is recommended.*

The use of a colored pencil in scoring is recommended because the number of correct answers on a page then stands out in bold relief and may be more quickly and accurately summed and recorded at the bottom of each page. A uniform method of marking the questions reduces the uncertainty a teacher

may feel in recorrecting a paper at a later date, in going over an examination paper with a pupil, or in later making statistical studies concerning the difficulty and validity of each question. The following marks are suggested: a check mark (✓) for right, a cross (×) for wrong, and an O for those omitted when necessary in right-minus-wrong scoring, to keep the omitted items separate from those wrong.

17. *The use of scoring formulas is not ordinarily recommended, with the possible exception of a "right-minus-wrong" formula for scoring true-false questions.*

In scoring single-word-answer questions or completion questions, the common practice is to give a credit of one point for each correct word supplied by the pupil. In scoring the single-choice or recognition questions a credit of one point for each correct underlining is also the rule. Experiments have demonstrated that no advantage arises from any effort to counteract the guessing element in this type of question by using such a formula as "number right minus one-third or one-fourth the number wrong." In scoring true-false questions, however, it may be advisable to use the "right-minus-wrong" formula, because the guessing element is so strongly present in these questions, even though great ingenuity to prevent the possibility of guessing be exercised in their preparation.

18. *The vexatious problem of "weighting" the marks assigned different questions can best be solved by the avoidance of any "weighting" whatsoever.*

This recommendation is based on the general lack of knowledge concerning the proper procedure for determining "weights." Furthermore, the problem of weighting is much less important when new-type examinations are used, because the large number of questions greatly reduces the

significance or importance of any given question. Weighting should be resorted to, of course, only when one question is much more important than another; and since there is no way of determining in advance the relative importance of each question, the problem of weighting loses significance and can safely be abandoned.[1]

19. *The total scores for the examination papers should be distributed on graph paper and a key for converting total scores into letter grades derived, and put in the form of a table.*

This recommendation recognizes the necessity of making a clear-cut distinction between measuring the capacity of pupils in terms of scores on the one hand and rewarding their efforts in the form of grades on the other. The measurement of pupil capacity is a technical problem, just as the determination of the height of individuals or the rate of basal metabolism of individuals is a problem involving scientific measurement. The awarding of grades, however, is an administrative problem involving essentially the establishment of educational policies. It is an interpretation of particular scores made on the examination with reference to established practices, standards of grading, or norms.

Measurements of capacity cannot be absolute at present because there are no defined zero points for the various subjects and there are no definite units of capacity each equal to every other. Therefore measurements of capacity must necessarily be relative — the score made by the average student in a group being the best point of reference, the other

[1] This recommendation deals with the question of weighting individual questions. However, when the various parts of the examination are disproportionate in length, it may be desirable to weight the scores for the various parts differently in deriving a total score for the examination as a whole. No single rule can be laid down for such procedure. This whole question of "weighting" must usually be settled arbitrarily and can best be avoided by endeavoring to prepare an examination with each part containing approximately the same number of questions.

scores being so much above or below that average. By distributing the scores made in an examination on a graph, the relativity of the scores will at once become apparent and the extent of the differences in capacity between the various pupils will be revealed. How many pupils should be given a grade of "A" then becomes a question of educational policy, just as the matter of how many individuals in a given group are to be designated "tall" on the basis of height measurements depends upon consensus of opinion and agreement as to just what height measurement must be exceeded by a person to warrant the label "tall."

Suppose that the prevailing practice demands 10 per cent A's, 20 per cent B's, 40 per cent C's, 20 per cent D's, and 10 per cent F's, and suppose that the teacher has no good reasons for assuming that his particular pupils differ in capacity from the average run of pupils; then a key for translating scores into grades could be derived as follows:

Scores in Exam.	Per Cent Making Scores	Grade to Be Assigned
134 to 170	9.5	A
105 to 133	21.0	B
83 to 104	39.0	C
55 to 82	20.5	D
0 to 54	10.0	F

The writer has no intention, in this place, of dictating or even suggesting the standard of grading to be followed by teachers in carrying out the recommendation for distributing the examination scores and transmuting them into letter grades. That standard may be made severe, resulting in few A's and B's and in a great number of D's and F's; or it may be lenient, resulting in many A's and B's and few D's and F's; or it may approximate some form of the normal frequency curve such as is used in the above example. The

standard to be used should be dictated by the prevailing practice in a given school or by the instructions issued by supervisors, principals, heads of departments, deans, or other administrative officers.

It scarcely need be pointed out that the relativity of measurements of capacity makes unnecessary the scoring of examinations on a percentage basis. Such a percentage basis has little meaning, since an 80 per cent mark on one examination may represent the capacity of one of the least capable pupils, whereas a 55 per cent mark on another examination may typify the work of one of the most capable pupils. The lack of meaning attaching to any particular percentage score arises, of course, because of the difficulty, *if not the impossibility*, of preparing a series of examinations each one of which is equal in difficulty to every other one. Because of these facts, the best practice in using new-type examinations consists in counting the number of right answers. The total possible score may, therefore, be 58, 93, 125, or what not, depending upon the number of units in the examination. It is not good practice to convert the number of points earned by a pupil into a percentage of the total possible number of points.

20. *The examination should be mimeographed or printed so that each pupil will have a copy; the examinations when scored should not be returned to the pupils.*

The nature of the many responses made in the new-type examination makes it necessary to submit to each pupil a mimeographed or printed copy. These should be filled out by the pupils and handed in to the teacher, care being taken to prevent any stray or extra copies falling into the hands of the pupils. Some teachers have attempted to give such examinations orally, thus avoiding the necessity of mimeographing the examinations. This practice is not recommended,

for it is reported that pupils develop signaling systems especially for the true-false and single-choice questions.

It is generally believed that various student organizations collect and file for the benefit of their members copies of various examinations given in the several courses. This results, of course, in unfair competition, because members of such organizations are favored by the opportunity to coach one another to answer typical sample questions, many of which are used over and over again in given courses. The new-type examination is so lengthy that pupils will be unable in the time allowed to copy it and attempts to copy can be prevented by proper proctoring. Hence the teacher can prevent the possibility of specific coaching for his examinations by simply *retaining under lock and key* all used and unused examinations. It is true that pupils have a right to know the results of their work on an examination, and this can be provided for by giving out examination scores and grades and by announcing that any pupil should feel free to consult the teacher concerning his actual work on any examination. This procedure has worked well both in satisfying the pupils that their individual interests are being protected and in preventing student organizations from becoming examination-coaching bureaus.

21. *The prevention of coaching should also be accomplished by using equivalent duplicate forms for classes in the same subject taking the examination at different hours or on different days and by changing the examination questions from semester to semester.*

This recommendation is in harmony with prevailing opinion and is to be followed in those situations where the pupils might know that the same examination was being repeated without change or where pernicious coaching bureaus or coaching tutors infest the educational community. Readers

are referred to the article by W. S. Miller, listed in the Appendix, in which the same examination has apparently been used successfully for ten successive quarters, as an instance where duplicate forms have been unnecessary.

22. *The diagnostic significance of each question should be determined and a large file of valid questions should be built up in order constantly to improve the accuracy of examinations.*

This recommendation is idealistic in so far as it looks forward to the distant day when teachers will have on file a very large number of *diagnostic* questions to serve as a reservoir for the ready assembly of highly accurate examinations. This recommendation urges various teachers to build up a file of all questions used for a period of years, so that a reservoir of from 1500 to 2000 objective questions could be developed from which examinations in endless variety could be quickly assembled and used as occasion demanded. Some provision would have to be made, of course, to eliminate from time to time those questions which become obsolete as the subject matter changes.

To carry out this recommendation completely would involve an enormous amount of clerical labor. The required methodology will be briefly described in terms of the work now actually under way by the Minnesota Department of Psychology for its own courses. As soon as a course is completed and the final course marks are reported to the registrar, a tabulation of the results for each objective question given during the course is made in the following manner:

a. Examination papers submitted by all students receiving final grades of A, B, D, or F in the course are segregated, together with a random sampling of at least 100 papers of students receiving the grade of C in the course.

b. The percentage of students in each of these five groups

who pass each question in the objective examination is then computed.

c. A card is made out for each question, giving the question and the acceptable answer, the date when the question was used, and the percentage of the A students passing it, the percentage of the B students passing it, and so on.

d. These cards are classified according to the form or type of question and the topic dealt with and are then filed for reference. An index of topics in the elementary course has been prepared for this purpose.

e. Whenever a question has been used a second time, its card is removed from the file and the percentage of A students passing it and so forth is recorded again, with a notation as to the date when it was used this second time.

The department, in following this method, is gradually building up a file of questions, each having been analyzed to determine the extent to which it differentiates between the superior, the average, and the inferior students in the course. The analysis for certain typical questions is given below as an illustration of the results.

Question: By the . . . . . (11) method we investigate likenesses and differences between individuals and groups.

This is a completion type of question, the correct answer being "comparative." The tabulated results for this question were:

	Course Marks				
	F	D	C	B	A
Per cent passing . . .	25	52	66	75	88

These figures show that there is a progressive increase in the per cent passing this question, proceeding upward from the F students to the A students. This is a good question,

for success in answering it is shown to have a close relationship to ability in the course as a whole.

Question: The junction of one neurone with another is known as a . . . . . (7).

This is a one-word-answer question requiring the correct answer "synapse" to be written over the seven dots. The results for the question were:

	Course Marks				
	F	D	C	B	A
Per cent passing . . .	90	100	100	100	100

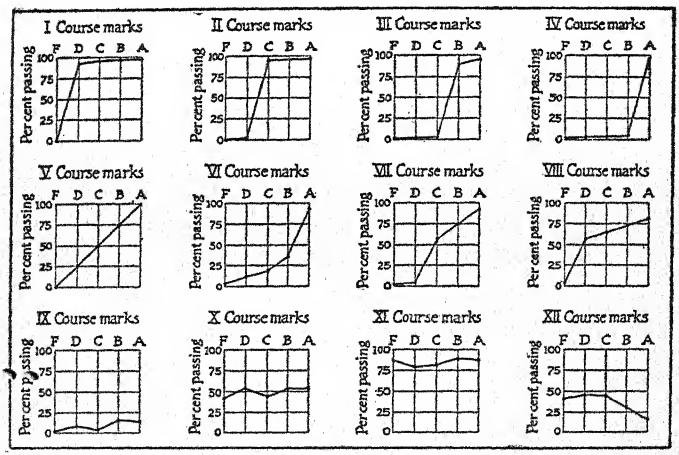
Here we have a question that is much too easy — so easy, in fact, that those of little achievement in the course (the F and D students) are able to pass it without difficulty. It is obvious that such a question neither adds nor detracts from the value of the examination, its presence merely adding to the total expense without serving any diagnostic purpose. However, such a question is useful if it is placed at the very beginning of an examination, with other equally easy questions, to serve as a "shock absorber."

Question: T F Unless a person intends to learn he cannot memorize.

This is a true-false question requiring T or F to be encircled according to the truth or falsity of the statement. The statement is actually false, and F should be encircled. The results for the question were:

	Course Marks				
	F	D	C	B	A
Per cent passing . . .	36	45	44	27	28

This question fails to differentiate between the five groups of students. Indeed, the A and B students actually show less ability in answering this question than do the F and D students. While it is true that this is an extreme example of an unsatisfactory question, yet such questions are rather frequent. In fact, the type of question given in the first illustration is rather rare and is far less common than one would ordinarily suppose in the absence of actual experience in attempting to discover experimentally the validity of examination questions. Additional examples of the results of such analytical work are given in the accompanying graph.



Graph illustrating different types of diagnostic questions. Small Charts I–IV illustrate ideal questions which sharply differentiate between one level of ability and all higher levels of ability. Such questions are so "ideal" that they are rarely found to exist in practice. Charts V–VIII illustrate satisfactory questions which show more gradual increases in the percentage passing them as you proceed from the F pupils to the A pupils. Charts IX–XI illustrate unsatisfactory questions which show little or no differences between the various levels of ability. Chart XII also illustrates an unsatisfactory question; in this case a smaller percentage of A and B pupils passed the question than of C, D, or F pupils. See J. C. Chapman's *Trade Tests* for a more complete discussion of similar analyses made in the construction of Army trade tests.

Some will question the wisdom of such a procedure in the absence of actual evidence demonstrating that diagnostic questions at one time continue to be diagnostic when used at a later time, and that non-diagnostic questions at one time continue to be non-diagnostic when used again. Unfortunately, such evidence is not plentiful, largely because this method of analysis has been used so infrequently. Our own departmental analysis has not been carried far enough to give adequate evidence on this point, but such results as we have indicate that a question found to be diagnostic tends to be diagnostic when used again. Only one published experimental study on this specific point has come to the writer's attention, that study concluding: "The stability of difficulty and distributing value of information questions is sufficiently stable so that questions can be evaluated one year for use the next." [1]

SUMMARY OF ABOVE DIRECTIONS

Perhaps a summary of the above directions in condensed outline form will be serviceable for ready reference purposes.

1. Questions covering every phase of the course should be utilized to insure wide sampling of pupil knowledge.

2. An excess number of questions should be prepared to allow ample opportunity for the selection of the best questions for the examination proper.

3. Ambiguous questions both with respect to meaning and possible answer should be rejected.

4. The apparent difficulty of a question should not be the basis for either accepting or rejecting a proposed question.

5. Acceptable questions should include an equal number of easy, hard, and moderately difficult questions.

[1] This conclusion is taken from the following article: W. R. Wilson, G. Welsh, and H. Gulliksen, "An Evaluation of Some Information Questions," *Journal of Applied Psychology*, Vol. VIII, No. 2, pages 206–214; June, 1924.

6. The first half dozen or so questions should be so easy that practically all can answer them, thus serving as a "shock absorber."

7. Each acceptable question should be an independent unit in the examination.

8. Each acceptable question should be short.

9. The examination should include a very large number of questions.

10. Each form or type of question should be segregated, the examination consisting of as many parts as there are types of questions.

11. Within each part of the examination the questions should be arranged according to topical sequence in the course.

12. The examination itself should be preceded by suitable general directions.

13. Specific directions should be given for each segregated group of questions.

14. There should be a random arrangement of true-false questions, with approximately an equal number of true and false statements.

15. The correct answers among the alternative answers in the single-choice and in the plural-choice questions should be placed according to chance.

16. A uniform method of marking the papers, together with the use of a colored pencil in scoring, should be used.

17. Scoring formulæ should not be used except possibly for the true-false type of questions, when a right-minus-wrong scoring formula may be used.

18. "Weighting" of questions according to difficulty or importance is rendered unnecessary in new-type examinations.

19. Total scores should be computed for the examination papers, distributed on a graph or table, and then a key for converting total scores into letter grades derived.

20. The examination should be mimeographed or printed, and both used and unused copies should be kept under lock and key in order to avoid the possibility of coaching.

21. The prevention of coaching should also be accomplished by using duplicate forms for classes in the same subject taking the examination at different hours or on different days and by changing the examination questions from semester to semester.

22. A large file of questions should be developed for each course, so that a reservoir of from 1500 to 2000 objective questions would be available from which examinations in endless variety could be quickly assembled and used as occasion demands. The ideal plan is to determine the diagnostic significance of each question, thus developing a large list of valid questions to be used in the preparation of examinations.

APPENDIX A

Illustrating the New-Type Examining Technique in Elementary Psychology [1]

This Appendix has been prepared in the belief that a presentation of the data yielded by an objective examination will be welcomed by those who have not experimented with such methods. It includes sample questions taken from the two-hour final examination given to one section in elementary psychology on December 16, 1924. These are followed by statistical data typifying certain profitable and illuminating manipulations of the examination results. The analyses here described are applicable in a variety of courses, whether taught in the elementary grades, in high school, or in college.

1. SAMPLE QUESTIONS

Part I consisted of 35 single-choice questions, the instructions being to underline the *one* word in parentheses which makes the *truest* or *best* statement. Ten of the questions were:

1. The appearance of a response after a succession of weak stimuli is called: (coördination; serial combination; summation; inhibition).
2. Allied reflexes (are learned; inhibit each other; reinforce each other; do not exist).
3. The autonomic nervous system (makes connections with smooth muscles; never acts except in emotion; has no connections with the central nervous system; is primarily a sensory system).
4. Glandular responses inhibited in emotional experiences are the (adrenal; thyroid; pituitary; salivary).
5. Euphoria is (a "higher emotion"; a primary emotion; an organic state; a state of extreme depression).
6. Hate may be considered as (an instinct; a primary emotion; a compound of anger and fear; a sentiment).

[1] The emphasis on examinations in this course results from the conditions under which it is given. No reliable class impression of the work of individual students is possible, since the quiz instructors meet each group only one hour a week for recitation. No laboratory work, no board work, no themes, no library reports are involved. Enrollment in the course is not open to freshmen; hence the number failing the course is much smaller than in a freshman course.

7. "Showing the teeth in scorn" is (an emotional state; a sensation; an expressive movement; a conditioned emotion).
8. Negative adaptation occurs quickly when the stimulus is (painful; very intense; natural; harmless).
9. In maze learning a rat is guided chiefly by (vision; hearing; smell; muscle-sense).
10. The jerky movements of the eyes in reading a book (are due to the nervousness of the person; can be suppressed at will; cannot be avoided; disappear with learning).

Part II consisted of 36 analogies. The instructions called for underlining the right word in the parentheses. Ten of the questions were:

1. Comparative method : differences and similarities : : genetic method : (development; degeneration; complete mastery; mnemonic systems).
2. Central fissure : frontal lobe and parietal lobe : : fissure of Sylvius : (parietal lobe and occipital lobe; parietal lobe and temporal lobe; occipital lobe and cerebellum).
3. Inhibition : coördination : : facilitation : (reinforcement; coördination; allied reflexes; mass action of cerebrum).
4. Spinal reflexes : cord : : natural balance movement : (cortex; thalamus; cerebellum; somesthetic area).
5. Knee jerk : lower spinal center : : speaking : (motor area; somesthetic area; super-motor area; thalamus).
6. Simple reaction (as in the reaction experiment) : reflex : : quick : (slow; quicker; hesitant; fast).
7. Native : acquired : : pupillary reflex : (sneezing; smiling; coughing; reading).
8. Anger : fighting : : fear : (rage; attack; elation; flight).
9. Play responses : fighting : : responses to organic needs : (locomotion; vocalization; self-assertion; sleep).
10. Removal of support : fear : : restraint : (disgust; anger; joy; sorrow).

Part III consisted of 25 completion statements which included a total of 85 blanks, the students being instructed to fill in the blanks with the correct word or words. Three of the completion statements were:

1. Behaviorism is . . . . . (8) . . . . . (8) psychology and regards psychology as a branch of . . . . . (7).

2. Within the neurone itself, functions are specialized as follows: the . . . . . (9) are the receiving organs, the . . . . (4) is the principal conducting organ, the cell-body is mainly . . . . . (9) in function, and the . . . (3) . . . . . (5) is the organ which transmits the impulse to the next nerve cell.
3. Detachment of a response occurs quickly when the response brings . . . . (4) and slowly when it brings . . . . . (7) in reaching the consummatory reaction.

The correct answers to the above sample questions were:

Part I	*Part II*	*Part III*
1. summation	1. development	1*a*. stimulus
2. reinforce each other	2. parietal lobe and temporal lobe	1*b*. response
3. makes connections with smooth muscles	3. coördination	1*c*. biology
4. salivary	4. cerebellum	2*a*. dendrites
5. organic state	5. super-motor area	2*b*. axon
6. compound of anger and fear	6. quicker	2*c*. nutritive
7. an expressive movement	7. reading	2*d*. end
8. harmless	8. flight	2*e*. brush
9. muscle-sense	9. sleep	3*a*. pain
10. cannot be avoided	10. anger	3*b*. failure

2. SCORING

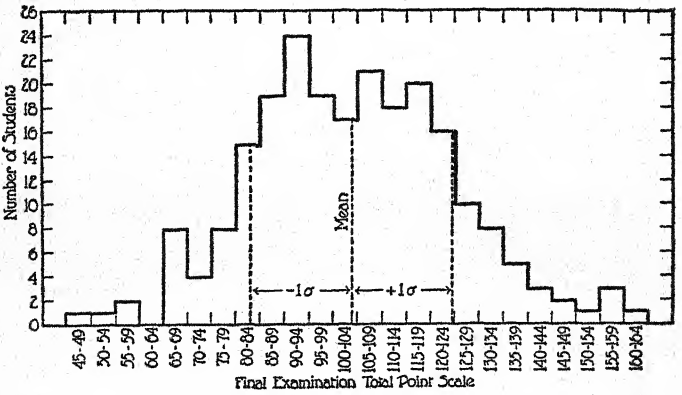
Correctly marked copies were prepared by the instructors and used as keys in scoring the papers. One point credit was given for each acceptable answer. The analogies were weighted by multiplying by two, giving in all a total possible score of 192. The scheme in table form is as follows:

Type of Question	Number of Possible Points	Weight	Number of Possible Weighted Points
Single Choice . .	35	1	35
Analogies	36	2	72
Completion . . .	85	1	85
		Total Possible	192

3. DISTRIBUTING THE TOTAL SCORES OR MARKS

As soon as total scores had been secured for each student's examination paper, they were plotted on a large sheet of co-ordinate paper. A scale was drawn on this sheet, allowing one coördinate distance for each point score in the examination, a small cross mark to indicate the score of any student being made above the appropriate point on the scale. Such a lengthy scale is used in order to facilitate the drawing of dividing lines between the A's and B's, etc. The accompanying graph shows the dis-

Frequency Distribution of Total Scores in Objective Final Examination in Elementary Psychology 1f, III Hour Section, 1924
(226 Students)



tribution of the scores made by the 226 students in this section. This graph differs from the original distribution sheet with respect to the scale intervals, there being a five-point range of scores on the base or scale line instead of one-point score intervals, and the number of students scoring at any particular point being represented by the height of the column instead of by a number of crosses one above the other, each representing the score of one student.

Attention is invited to the fact that these examination scores approximate the "normal probability curve," the great mass of scores piling up in the middle ranges of the scale, whereas scores

deviating in either direction from the middle become less and less frequent. The average score is 104.0 points, with approximately 68 per cent of the scores falling between 83 and 125.

4. KEY FOR CONVERTING SCORES INTO LETTER GRADES

The next step in the treatment of the scores involved the determination of the score limits to be used in transmuting the crude scores to letter grades. These limiting points were determined by the instructors in charge of the course, in part by inspection of the nature of the distribution of scores and in part by determination of various score limits, including certain percentages of the students. This done, lines were drawn on the original distribution sheet separating the A's from the B's, the B's from the C's, etc. Then the following key for converting crude scores into letter grades was prepared:

KEY FOR CONVERTING SCORES INTO LETTER GRADES[1]

SCORE RANGE	LETTER GRADE	NUMBER	PER CENT
127–228 (inclusive) . . .	A	27	11.9
118–126 "	B	32	14.1
90–117 "	C	109	48.2
83–89 "	D	28	12.4
0–82 "	F	30	13.3
	Total	226	99.9

It will be noted that the proportion of A's and F's practically equals the proportion of B's and D's, and that these percentages are slightly larger than those contemplated as final course grades. This is done for the purpose of counteracting in part the inevitable restriction of range of average grades when the several grades for each student are averaged in determining the final letter grades for the course as a whole.

[1] In actual practice the department usually further subdivides the C group into C+, C, and C−, a small percentage of the papers being given a C+ grade or a C− grade. These qualified C grades are helpful in averaging a number of grades when a student's average grade is on the borderline between a B and a C or between a C and a D.

5. DETERMINATION OF COURSE MARK[1]

Each instructor proceeds to post in his grade book the actual scores and the corresponding letter grades on the final examination. He then averages the grades for the ten weekly written quizzes (each ten minutes in length) given during the quarter, by giving numerical values of 4, 3, 2, 1, or 0 to the grades of A, B, C, D, or F, respectively, and then averaging these numerical values. The letter grades in the mid-quarter and the final examinations are given numerical values in the same way. The average grade on the weekly quizzes, the grade on the mid-quarter examination, and the grade on the final examination are then pooled to form the final grade in the course. Departmental custom decrees that the weekly quizzes shall be given a weight of three sixths, the mid-quarter examination a weight of one sixth, and the final examination a weight of two sixths; hence it is necessary to multiply the weekly quiz average by 3, the mid-quarter examination by 1, and the final examination by 2, and then to divide the sum of these three grades so weighted by 6. The distribution of grades in each of these three components of the final course mark, together with the distribution of the final course marks themselves, is as follows:

	Average of Weekly Quizzes	Mid-quarter Grade	Final Examination Grade	Final Course Mark
A	20	17	27	19
B	34	31	32	39
C	116	115	109	108
D	35	33	28	41
F	21	30	30	19
Total	226	226	226	226

[1] The three-hour course in elementary psychology is conducted as follows: Two lectures a week are given to all 226 students in one group; the 226 students then meet for a third hour in small sections for quiz, review, and discussion of the lectures and assignments. At the beginning of each quiz hour a ten-minute written quiz is given to each section. The mid-quarter and final examinations are given to all 226 students in one group, being uniform for all.

6. RELATION OF VARIOUS COMPONENTS TO EACH OTHER AND TO FINAL COURSE MARK

The fact that the distributions given in Section 5 are similar is no guarantee that the grades given to the individual students in any one component will be closely related to those given the same individuals in the other components. Hence it is desirable to know the extent of the relationship existing between these various examinations for estimating each individual's performance in the course. These relationships are revealed in the following "scatter tables":

1. RELATION BETWEEN GRADES IN FINAL EXAMINATION AND COURSE MARK

FINAL EXAM. GRADES	COURSE MARK					TOTAL	PER CENT
	F	D	C	B	A		
A			1	9	17	27	12.0
B			9	21	2	32	14.1
C		16	84	9		109	48.3
D	1	14	13			28	12.4
F	18	11	1			30	13.2
Total	19	41	108	39	19	226	100.0
Per Cent	8.4	18.1	47.8	17.2	8.4	99.9	

Pearson product-moment coefficient of correlation $= + .855 \pm .012$

Summary of Agreements or Disagreements in Above Table:

	No.	PER CENT
Perfect agreement[1] in	154	68.1
Disagreement of one step in	70	31.0
" " two steps in	2	.9
Total	226	100.0

The above relationship is remarkable in spite of the fact that it must be discounted slightly because two sixths of the course mark itself is determined by the final examination grade. It is especially significant to find 68.1 per cent of the cases in perfect agreement, with not a single disagreement of three or four steps.

[1] Perfect agreement indicates identity between the mark in the examination and the mark in the course.

2. Relation between Grades in Mid-quarter Examination and Course Mark

Mid-quarter Examination Grades	Course Mark					Total	Per Cent
	F	D	C	B	A		
A				3	14	17	7.5
B			5	22	4	31	13.7
C	1	12	87	14	1	115	50.9
D	2	19	12			33	14.6
F	16	10	4			30	13.3
Total	19	41	108	39	19	226	100.0
Per Cent	8.4	18.1	47.8	17.2	8.4	99.9	

Pearson product-moment coefficient of correlation $= + .83 \pm .014$

Summary of Agreements or Disagreements in Above Table:

	No.	Per Cent
Perfect agreement	158	70.0
Disagreement of one step	62	27.4
" " two steps	6	2.6
Total	226	100.0

Here again the correlation is remarkably close, even though the mid-quarter examination contributes only one sixth to the course work. There is even a slightly larger number of perfect agreements, with no serious three- or four-step disagreements.

3. Relation between Average Weekly Quiz Grades and Course Mark

Average Weekly Quiz Grade	Course Mark					Total	Per Cent
	F	D	C	B	A		
A				5	15	20	8.9
B			12	18	4	34	15.0
C	2	14	84	16		116	51.4
D	6	19	10			35	15.5
F	11	8	2			21	9.2
Total	19	41	108	39	19	226	100.0
Per Cent	8.4	18.1	47.8	17.2	8.2	99.9	

Pearson product-moment coefficient of correlation $= + .80 \pm .016$

Summary of Agreements and Disagreements in Above Table:

	No.	Per Cent
Perfect agreement	147	65.0
Disagreements of one step	75	33.2
Disagreements of two steps	4	1.8
Total	226	100.0

Once again a close relationship is demonstrated, although greater discount needs to be made because three sixths or one half of the course mark itself is dependent upon the average weekly quiz grade. There are no disagreements of a three- or four-step magnitude.

4. Relation between Mid-quarter Examination Grades and Final Examination Grades

Mid-quarter Examination Grades	Final Examination Grades					Total	Per Cent
	F	D	C	B	A		
A				2	15	17	7.5
B			10	13	8	31	13.7
C	4	14	77	16	4	115	50.9
D	8	10	14	1		33	14.6
F	18	4	8			30	13.3
Total	30	28	109	32	27	226	100.0
Per Cent	13.2	12.4	48.3	14.1	12.0	100.0	

Pearson product-moment coefficient of correlation $= + .74 \pm .02$

Summary of Agreements and Disagreements in Above Table:

	No.	Per Cent
Perfect agreement	133	58.9
Disagreement of one step	76	33.6
" " two steps	17	7.5
Total	226	100.0

It must be remembered that we are comparing here the status of the students in each of two objective examinations covering for the most part different subject matter and separated by a time interval of seven or eight weeks. This would seem to mean that both examinations are measuring consistently a capacity to learn

the content and profit from the instruction in elementary psychology and, further, that this capacity is a characteristic manifested by these students in approximately equal relative amounts at both the middle and the end of the quarter. According to accepted standards, current in the field of diagnostic testing of human traits, this is evidence of the high reliability of this method of examination.

5. Relation between Average Weekly Quiz Grades and Final Examination Grades

Average Weekly Quiz Grade	Final Examination Grades					Total	Per Cent
	F	D	C	B	A		
A			2	5	13	20	8.9
B		3	15	12	4	34	15.0
C	10	18	66	12	10	116	51.4
D	10	3	19	3		35	15.5
F	10	4	7			21	9.2
Total	30	28	109	32	27	226	100.0
Per Cent	13.2	12.4	48.3	14.1	12.0	100.0	

Pearson product-moment coefficient of correlation $= + .57 \pm .03$

Summary of Agreements and Disagreements in Above Table:

	No.	Per Cent
Perfect agreement	104	46.0
Disagreement of one step	87	38.5
" " two steps	35	15.5
Total	226	100.0

The above relationship is the lowest of those reported in this Appendix, although even here we find agreement within one step in the case of 84.5 per cent of the students. This lower relationship would be interpreted by some as due to the fact that weekly quizzes are measuring a different phase or aspect of the student's capacity in psychology from that which the final examination measures; hence they would conclude that it is wise to have the course mark reflect both kinds of achievement as representative of what the student can do in the course as a whole.

There are reasons, however, for believing that such an interpretation errs in the direction of overvaluing the reliability and significance of the average weekly quiz grade.

The following reasons make us hesitate to place such reliance on the weekly essay-type quizzes. In the first place, the 226 students were divided among eight quiz sections, their weekly quizzes being read by three graduate assistants whose grading cannot be considered to be infallible, and whose standards of grading must inevitably vary among themselves and are likely to vary from week to week. In the second place, the questions constituting the weekly quizzes were not uniform for all sections in any given week, such lack of uniformity involving undesirable but uncontrollable variations in the difficulty of the questions. In the third place, only one or two general questions of the essay type could be used each week for any given section; hence the questions inadequately sample the content covered in the two preceding lectures and the forty to sixty page assignments. These constant and variable errors are supplemented by others of less importance not mentioned. All together, one would be safe in concluding that the average of even ten weekly quizzes must fall short of the standards of good examining technique.

For these reasons the writer would be inclined to place much more emphasis on the objective mid-quarter and final examinations and would even go so far as to eliminate the weekly quiz grades from the course mark, were it not for the incentive to consistent weekly application that is afforded by the weekly quizzes and the knowledge that they are to count one half toward the course mark. Realization of the errors involved in the present weekly essay-type quizzes and at the same time realization of the importance of these quizzes as an incentive to consistent application has led the staff to begin the accumulation of short-answer questions for use in weekly quizzing. These are being tried out in four sections of the course this year, access to a "Ditto" duplicating machine making possible the plan of using a one- or two-page new-type examination of from 30 to 40 questions to be answered in ten minutes.

7. SUMMARY OF APPENDIX A

It is hoped that the preceding sections may serve as illustrative material descriptive of the various steps in the scoring of the objective examinations, graphing the examination scores, conversion of the crude scores into letter grades, determining the final course mark, and analyzing statistically the relation between the marks derived from the various components of the course mark. It is the hope of the writer that other instructors may be sufficiently interested in this technique to analyze their own course marks in a similar manner. He feels certain, from a knowledge of like analyses in courses using traditional examinations, that few courses will be found where the various components will show such close relationships with final course marks or with each other. In other words, he predicts that teachers adopting this analytical method will find many more inconsistencies and will be led, naturally, to take a more questioning attitude toward their grading problems and to experiment with more uniform and objective methods of measuring achievement.

APPENDIX B

Annotated Bibliography on New-Type Examinations

While an effort has been made to prepare as complete a bibliography as possible, yet no claim for completeness is made. No references to the voluminous work on standardized achievement tests or standard intelligence tests are included. References preceded by an asterisk are considered of unusual importance.

Asker, William. "The Reliability of Tests Requiring Alternative Responses." *Journal of Educational Research*, Vol. 9, No. 3, pages 234–241; March, 1924.

Analyzes the two-alternative (true-false type) and the three-alternative (recognition type) tests on the basis of experiment and mathematics and finds that a completely ignorant person has a very small chance of obtaining a passing mark. Urges discouraging of guessing and recognizes the need for a more reliable method of scoring such tests.

Barthelmess, H. M. "Reply to Criticism of Tests Requiring Alternative Responses." *Journal of Educational Research*, Vol. 6, No. 4, pages 357–359; November, 1922.

Denies the validity of criticisms directed against true-false tests and affirms that they do have genuine pedagogical value.

Batson, W. H. "Reliability of the True-False Form of Examination." *Educational Administration and Supervision*, Vol. 10, pages 95–103; 1924.

Describes an experiment using four true-false examinations and four essay examinations in a course in elementary education, concluding, "Results obtained by the True-False Examination conform sufficiently to those obtained by the regular examination to make it possible for the True-False Examination to be substituted for the essay examination."

Blumer, G. "Desirability of Changing the Type of Written Examinations." *Journal of American Medical Association*, Vol. 72, pages 1131–1133; 1919.

Argues against the traditional medical examinations as mere memory tests and proposes to adopt objective technique, developed by the psychologists, in the preparation of medical examinations.

Chapman, J. C. "Individual Injustice and Guessing in the True-False Examination." *Journal of Applied Psychology*, Vol. 6, pages 342–348; 1922.

Another protest against assuming that a right-minus-wrong scoring formula for true-false tests counteracts the effects of guessing and results in accurate scores for each individual. "With other forms of examination, such as the single-word-answer or completion type, the evils which accompany guessing do not enter."

CHAPMAN, J. C. "The Measurement of Physics Information." *School Review*, Vol. 27, pages 748–756; 1919.

Describes the development and trial of a thirty-question objective test in high school physics utilizing the one-word-answer method.

*CHAPMAN, J. C. *Trade Tests.* Henry Holt & Co., New York; 1921. ix + 435 pages.

A detailed account of the work of the United States Army Trade Test Division of the Committee on Classification of Personnel in the Army. This book is especially valuable for the excellent presentation of the principles underlying standardized trade-test technique in which the traditional multi-answer questions are contrasted with the new single-answer questions.

CHAPMAN, J. C., and TOOPS, H. A. "A Written Trade Test: Multiple-Choice Method." *Journal of Applied Psychology*, Vol. 3, No. 4, pages 358–365; 1919.

An experimental demonstration of the adaptation of the oral trade-test method to the examination of a group of individuals by means of a printed trade test, the answers being written on the examination sheet.

COLVIN, S. S. "Marks and the Marking System as an Incentive to Study." *Education*, No. 32, pages 560–572; May, 1912.

A theoretical discussion of the inaccuracy of school marks and the urgent need of making them accurate in order that their uses as an incentive to study may be permitted full scope. Advocates the development and use of objective tests of school achievement in order to accomplish this aim.

COOLEY, A. M., and REEVES, G. "Some Investigations Concerning the Use of Certain Home-Economics Information Tests." *Teachers College Record*, Vol. 24, No. 4, pages 374–392.

Illustrates the use of the multiple-choice or recognition type of question in preparing extensive and comprehensive information examinations in the field of home economics.

DALMAN, M. A. "Hurdles, a Series of Calibrated Objective Tests in First-Year Algebra." *Journal of Educational Research*, Vol. 1, No. 1, pages 47–62; 1920.

Describes a series of objective tests in algebra, with certain statistical evidence showing their usefulness.

DICKINSON, Z. C. "Suggestions toward Improving Examination Marks." Report of the Committee on Educational Guidance, Bulletin of the University of Minnesota, Vol. 26, No. 31, pages 31–36; August 4, 1923.

A discussion of the usefulness of new-type examinations as a partial basis for grading college students.

*FILER, H. A., and O'ROURKE, L. J. Annual Reports of the Chief Examiner and the Director of Research of the United States Civil Service Commission for the Fiscal Year Ended June 30, 1923, pages 1–64. Government Printing Office, Washington, D. C.; 1923.

A detailed description, accompanied by statistical analyses, of the extensive new-type examination experiments being conducted by the United States Civil Service Commission. See other annual reports of the Chief Examiner since 1919 for further information on this work.

FISKE, T. S. Annual Reports of the Secretary of the College Entrance Examination Board, 431 West 117th Street, New York; 1921, 1922, 1923, and 1924.

Brief reports are contained in these annual reports of the various developments leading the C. E. E. B. to adopt, in part, new-type examination technique in certain subjects.

GATES, A. I. "The True-False Test as a Measure of Achievement in College Courses." *Journal of Educational Psychology*, Vol. 12, pages 276–287; 1921.

Presents much correlational data obtained from ten classes in educational psychology, Columbia University, concerning the true-false examination in relation to essay examinations, concluding that the evidence more than justifies its use in examination procedures.

GRAY, WILLIAM S. "Value of Informal Tests of Reading Accomplishment." *Journal of Educational Research*, Vol. 1, No. 2, pages 103–111; 1920.

Stresses the importance of informal tests prepared by the teachers, holding that the existence of standardized objective reading scales does not make the informal test unnecessary.

HAHN, H. H. "A Criticism of Tests Requiring Alternative Responses." *Journal of Educational Research*, Vol. 6, pages 236–240; October, 1922.

Attacks the accuracy of true-false tests and the right-minus-wrong formula for scoring them, and also holds that they are undesirable from various pedagogical points of view.

HAYES, SETH. "Coöperative Chemistry Tests." *Journal of Educational Research*, Vol. 4, No. 2, pages 109–120; 1921.

Describes the coöperative development and use of eight hundred objective questions based on McPherson and Henderson's textbook, in the high schools of Cleveland, Ohio.

JACKSON, DUNHAM. Letter to the Editor of the *Harvard Alumni Bulletin* in April, 1922.

Criticizes Wood's reliability technique in his College Entrance Examination Board investigation, pointing out that correlations of halves of a test should be low if the various questions are not designed to measure the same thing but to measure different aspects of the candidate's reactions to the instruction that he has received.

JAMES, B. B. "The Modern Test." *School and Society*, Vol. 19, pages 209–213; February, 1924.

Describes the new-type examination and gives seven reasons why it is better than the topical answer.

KNIGHT, F. B. "Data on the True-False Test as a Device for College Examination." *Journal of Educational Psychology*, Vol. 13, No. 2, pages 75–80; February, 1922.

Reports an experiment with the true-false examination in elementary physics at the State University of Iowa, concluding that "a more thorough true-false test including ingenious statements concerning laboratory technique can be expected to do as well, if not better, than written examinations, with the sound advantage clearly on its side of saving the instructor's time."

KOHS, S. C. "High Test Scores Attained by Subaverage Minds." *Psychological Bulletin*, Vol. 17, pages 1–5; 1920.

Attempts to formulate the mathematics of guessing as it pertains to the two-alternative or true-false type of test, assuming that responses to such a test are completely naïve and based upon mere chance alone. Shows that under such conditions with a fifty-item test the chances are small that any one would receive a score of more than 16 per cent correct. (Chance of such a score is 1 out of 17.)

LAIRD, DONALD A. "A Comparison of the Essay and the Objective Type of Examinations." *Journal of Educational Psychology*, Vol. 14, No. 2, pages 123–124; 1923.

Shows that the average student exhibits twice as much information about a topic when given an objective test as when given an essay test on the topic.

MCAFEE, L. O. "The Reliability of Non-Standardized Point Tests." *Elementary School Journal*, Vol. 24, No. 8, pages 579–585; April, 1924.

A comparison of three forms of new-type questions and the essay-type question given to fifty-seven seventh-grade children in American history. Concludes that new-type-question tests give more reliable results than the essay-question test and that one-word-answer questions are slightly more reliable than true-false questions.

McCall, W. A. *How to Measure in Education.* The Macmillan Company, New York; 1922.

This book attempts to present the essentials of educational measurements and includes a discussion of the new-type content examinations.

McCall, W. A. "A New Kind of School Examination." *Journal of Educational Research,* Vol 1, page 33; 1920.

Describes the true-false form of question, discusses methods of scoring, gives instructions for their preparation, and claims certain advantages for this form of examination which make it in his opinion "a herald of newer and better types of examination." Points out that 12,000,000 examinations are given in the schools each year; hence the importance of examination technique.

* May, M. A. "Measuring Achievement in Elementary Psychology and in Other College Subjects." *School and Society,* Vol. 17, No. 435, pages 472–476 (April 28, 1923); Vol. 17, No. 438, pages 556–560 (May 19, 1923).

Presents statistical data and norms for a standardized new-type examination covering the subject matter in Woodworth's *Psychology.*

May, M. A. "Standardized Examinations in Psychology and Logic." *School and Society,* Vol. 11, pages 533–540; May 1, 1920.

Gives sample objective questions of different types used in psychology and logic, together with a statistical technique for treating the examination scores.

Miller, G. F. "A Variation in the 'True and False' Achievement Test." *School and Society,* Vol. 20, No. 504, page 250; August 23, 1924.

Describes a true-false test in educational psychology in which the student is to assign the numbers 1 to those statements which are true, 2 to those which are false, and 3 to those which may be either true or false.

* Milner, W. S. "An Objective Test in Educational Psychology." *Journal of Educational Psychology,* Vol. 16, No. 4; April, 1925.

Presents statistical evidence of the reliability and validity of a new-type examination in educational psychology which was used for ten successive quarters without significant change either in the measures of central tendency or of variability.

Monroe, W. S. "Written Examinations and Their Improvement." Bulletin No. 9 of the Bureau of Educational Research, University of Illinois.

Monroe, W. S. "Written Examinations versus Standardized Tests." *School Review,* Vol. 32, No. 4, pages 253–265; April, 1924.

A comparison of the reliability of written examinations and standardized tests. Believes "that teachers can make material reductions in

both constant and variable errors in examination grades if they observe certain rules in the preparation and administration of written examinations."

* MONROE, W. S., and SOUDERS, L. B. "Present Status of Written Examinations and Suggestions for Their Improvement." Bulletin No. 17 of the Bureau of Educational Research, University of Illinois; 1923.

Reports of details of an extensive investigation of present-day tendencies in giving written examinations in high schools, including data on the reliability of actual written examinations and instructions for supplementing the traditional examinations by the adoption of the new-type-examination technique.

ODELL, C. W. "Another Criticism of Tests Requiring Alternative Responses." *Journal of Educational Research*, Vol. 7, pages 326–330; April, 1923.

Criticizes those who object to right-minus-wrong scoring devices for two-alternative responses, such as true-false tests, and gives arguments in favor of such scoring methods. Also points out the value of such tests in examining procedures.

* PATERSON, D. G. "Improving the Examination Function in Teaching." Report of the Committee on Educational Guidance, Bulletin of the University of Minnesota, Vol. 26, No. 31, pages 47–56; August 4, 1923.

A discussion of the claims for the new-type examinations, together with illustrations of different types of questions and some statistical evidence of the usefulness of these examinations in the field of psychology.

POWERS, S. R. "A Comparison of Achievement of High School and University Students in Certain Tasks in Chemistry." *Journal of Educational Research*, Vol. 6, pages 332–343; 1922.

This work illustrates the use of the multiple-choice or recognition method in measuring certain abilities in elementary chemistry in high schools, colleges, and universities.

REMMERS, H. H., MARSCHAT, L. E., BROWN, A., and CHAPMAN, I. "Experimental Study of the Relative Difficulty of True-False, Multiple-Choice, and Incomplete-Sentence Types of Examination Questions." *Journal of Educational Psychology*, Vol. 14, No. 6, pages 367–372; September, 1923.

A preliminary study on some results obtained in an effort to determine the relative difficulty of these three forms of new-type questions.

RICHARDS, O. W. "High Test Scores Attained by Subaverage Minds." *Journal of Experimental Psychology*, Vol. 7, pages 148–156; 1924.

Demonstrates by statistical theory the probable scores to be obtained on the basis of guessing alone in tests with two, three, and four alterna-

tive answers. Gives no consideration to the psychological factors in such tests which remove them from the realm of pure chance.

RICHARDS, O. W., and KOHS, S. C. "High Test Scores Attained by Sub-average Minds." *Journal of Educational Psychology*, Vol. 16, No. 1, pages 8–18; January, 1925.

Demonstrates by analysis of the laws of chance that extreme care must be used in interpreting results derived from true-false tests and recommends that a true-false test include at least seventy-five items to obviate the chance error in scores.

SANFORD, VERA. "A New-Type Final Geometry Examination." *Mathematics Teacher*, Vol. 18, No. 1, pages 22–37; January, 1925.

Describes the development of a new-type geometry examination in Lincoln School of Teachers College, together with statistical evidence of its reliability and validity.

* SEASHORE, C. E. "College Placement Examinations." *School and Society*, Vol. 20, No. 515, pages 575–580; November 8, 1924.

Describes the Iowa Placement Examinations in English, French, Mathematics, and Chemistry, which utilize the multiple-choice or recognition question, and advocates their experimental use in sectioning students into homogeneous groups on the basis of ability.

* SEASHORE, C. E. "Progressive Adjustment *vs.* Entrance Elimination in a State University." *School and Society*, Vol. 17, No. 420; January 13, 1923.

Advocates the use of objective examinations to make college marks more reliable, so that they may be used more safely in educational guidance.

TELFORD, FRED. "The Work of the Board of Examiners of the New York City Board of Education." Public Personnel Studies, Bureau of Public Personnel Administration, Vol. 2, No. 9, pages 268–287; December, 1924.

A detailed survey of the work of this board of examiners, indicating that the board has incorporated the short-answer form of question in its examining procedures. See other issues of Public Personnel Studies for instances of the use of new-type questions in civil service examinations.

* THORNDIKE, E. L. "The Future of the College Extrance Examination Board." *Educational Review*, Vol. 31, pages 470–479; 1906.

A statistical analysis of the weaknesses of traditional examination methods as used by the C. E. E. B., showing low correlations between entrance examinations and college marks. Thorndike concludes, "It is certain that the traditional entrance examinations, even when as fully safeguarded as in the case of those given by the College Entrance Examination Board, . . . do not measure fitness for college well

enough to earn the respect of students or teachers, and do intolerable injustice to individuals."

THORNDIKE, E. L. "Entrance Exams and College Grades." *Science*, New Series, Vol. 23, pages 839–845; 1906.

A more detailed presentation and analysis of the statistical data forming the basis of the preceding article.

* TOOPS, H. A. *Trade Tests in Education.* Contributions to Education, No. 115. Teachers College, Columbia University, New York; 1921. vi + 118 pages.

An excellent dissertation demonstrating the adaptation of the technique of trade-test construction to the preparation of school examinations.

* WEISS, A. P. "On Methods of Mental Measurement, Especially in School and College." *Journal of Educational Psychology*, Vol. 2, pages 555–563; 1911.

Presents a method of obtaining comparable scores from different tests, utilizing data obtained by the completion-test method, comprising two objective quizzes given to a section of the class in the Introduction to Psychology. This is a reference to some of the earliest experiments in the use of objective examination methods in college teaching.

WEST, P. V. "A Critical Study of the Right-Minus-Wrong Method." *Journal of Educational Research*, Vol. 8, pages 1–10; 1923.

Severely criticizes the right-minus-wrong scoring formula applied to true-false tests, because when he plotted scores obtained from a true-false test and scored right minus wrong, his resulting curve did not conform to the so-called normal curve of error.

WIGMORE, J. H. "The 'New Type' Law Examination." *Illinois Law Review*, Vol. 19, No. 3, pages 172–173; November, 1924.

A letter from Dean Wigmore, Northwestern University, to the editor, counting himself in favor of the new-type examination, having tried it in three subjects for two years, but taking issue with Ben D. Wood's recent article, "The Measurement of Law School Work," by criticizing the right-minus-wrong scoring formula for true-false questions because this formula holds only for averages and may in a given case be in error. Hypothetical data only are used as the basis of reasoning.

* WOOD, BEN D. *Measurement in Higher Education.* World Book Company, Yonkers-on-Hudson, New York; 1923. xi + 337 pages.

A detailed account of the Columbia College experiments with the Thorndike Intelligence Examination and the New-Type Content Examinations. This is probably the most important single reference on the use of new-type examinations in various college subjects.

* WOOD, BEN D. "The Measurement of Law School Work." *Columbia Law Review*, Vol. 24, No. 3; March, 1924. 42 pages.

Presents experimental evidence of the need for and the value of new-type content examinations in law schools.

WOOD, BEN D. "The Measurement of College Work." *Educational Administration and Supervision*, pages 301–331; September, 1921.

A detailed report of the results of the new-type examining technique as applied in contemporary civilization, Columbia College. The facts presented in this report appear again in Wood's book, *Measurement in Higher Education*.

* WOOD, BEN D. "The Reliability and Difficulty of the College Entrance Examination Board Examinations in Algebra and Geometry." College Entrance Examination Board, 431 West 117th Street, New York; 1920.

A detailed statistical report of an investigation concerning the reliability and difficulty of the C. E. E. B.'s traditional examining procedures in algebra and geometry, with recommendations for the partial adoption of new-type technique. This report led the C. E. E. B. to appoint a committee of its own to investigate the claims for the new-type examination.

* Report of Commission on New-Type Examinations. College Entrance Examination Board, 431 West 117th Street, New York; November 3, 1923.

A comparison of new-type and old-type examinations, with reference to reliability or internal consistency and validity or the extent to which each is correlated with preparatory-school marks and college marks. This report convinced the C. E. E. B. of the desirability and wisdom of incorporating new-type examining technique as a part of their procedures in certain subjects.

As this manual in revised form was being assembled for submission to the publisher, announcement came of the publication of an important contribution to this subject by G. M. Ruch, entitled *The Improvement of the Written Examination*, published by Scott, Foresman & Co., Chicago; 1924 (193 pages). This book discusses the functions of written examinations, the criteria of a good examination, sources of error in written examinations, types and construction of the newer objective examinations, with many examples of new-type questions in various subjects and experimental studies of the relative merits of the various forms of new-type examinations. The book includes data published by G. M. Ruch and G. D. Stoddard in the January-February, 1925, issue of the *Journal of Educational Psychology*, under the title "Comparative Reliabilities of Five Types of Objective Examinations."